Norwegian School of Economics

ESSAYS ON STAFFING AND TRANSPORTATION

Christian Braathen

Dissertation submitted to the Department of Business and Management Science, Norwegian School of Economics, in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD)

Bergen,
January 2024

Acknowledgments

"It takes a village to raise a child" is among my favorite proverbs. The idea is that a person can realize their potential to a greater degree if a whole community comes together to give the support, education, care, and wisdom that a young person needs to live an enriching life.

To assume that my achievements are mine and mine alone would be wildly ignorant. While Sir Isaac Newton's quote "If I have seen further, it is by standing on the shoulders of giants" is commonly attributed to scientific discoveries being built upon the foundational work of previous scholars, I believe the quote is also fitting in the context of the village. Because had the extensive, but far from exhaustive, list of people below not come together for me, then this directionless dropout would likely have remained just that—and definitely not have achieved a PhD. I am deeply grateful that each and every one of you has taken me on your shoulders.

Thank you, giants.

Mario Guajardo

First and foremost, I must give a massive thank you to my supervisor, Mario Guajardo. Little did I know that February 19th, 2015, would become the most important day of my academic life because of Mario. In one of my first-year courses, he blended my newfound excitement for mathematics and my long-lasting love for business and football when he gave a guest lecture on using optimization to create a dramatically improved match schedule for a football league.

Not only did Mario give a remarkable lecture showing that real problems are hard due to their combinatorial structure, large scale, and many criteria, but he also cheekily suggested that my favorite Norwegian football team would have won the league in 2007 if Tippeligaen had used an optimal schedule instead of the simplistic circular method they already followed. I still remember the awestruck gasp that went through the auditorium when Mario showed the list of considerations that had to be made in real-life problems. But he had already won me over—real-life optimization problems were the most intriguing academic problems I had seen.

After a great talk during the intermission of that lecture, I reached out to Mario again a year

later. I had discovered my own real-life optimization problem I wanted to solve in preference-based shift scheduling, and I lacked the tools and know-how to get started. Mario not only provided me with the tools and resources I needed but also helped me in five minutes with a modeling challenge I had not been able to solve for months. Those five minutes demonstrated how much I had yet to learn and inspired me to develop further with Mario's guidance. First by enrolling in his master's course in 2017, then with Mario as my master's thesis supervisor in 2018, and as my PhD supervisor from 2019 onwards.

During my PhD years, I have been enormously grateful for how Mario has guided, shared valuable wisdom and recommendations, and supported me as I explored various research questions. And most importantly—how he has given me the freedom to work on projects that have been academically and intellectually fun for somebody with a passion for real-life optimization problems.

There's only a handful of people you meet in your life that will dramatically change your life in a myriad of ways. Mario has been one of them. Had Mario not replied to the inquiry of a curious second-year bachelor student with a vague problem he would like to solve on his own, my life today would have been dramatically different. So *thank you* for being kind, showing support, and researching with passion for nearly a decade of knowing you.

Jan Ubøe

Before starting at NHH, I had never performed well in school—it was practically only the age points and prior studies that gave me the GPA to be accepted to NHH. After dropping out of two bachelor programs prior to NHH, I would never in my wildest dreams believe undertaking a PhD at Norway's best business school is something I could be capable of doing. Luckily, though, I met Jan Ubøe. Jan saw something in me that neither I nor anybody else did, and he gave me the chance to test my PhD capabilities during my second year of the bachelor's program. He gave me a chance to enroll in his PhD course "Topics in Probability Theory and Stochastic Processes", which challenged me more academically than I had ever been pushed at that point, but I also revelled the moments Jan enthusiastically declared "That's right!" when I got the mathematics and interpretations correct. Had it not been for Jan, I would never have applied to a PhD program, and I am deeply grateful and honored that we wrote an engaging and important research article together.

Jan Ubøe and Inge Thorsen

During my research project with Jan and Inge, I witnessed the greatest display of an interdisciplinary team that I have ever been part of myself. It was astonishing to see how three researchers with very different skill sets started out working on one research problem, realized an important pre-problem that had to be solved, and then spent two years on interpreting and deducing the nooks and crannies of that pre-problem instead—producing great enough results that we wanted to share it with the world as a research article. I am enormously grateful for all the time we spent together in solving this research question. But maybe more importantly, I am grateful for how we worked as a team.

Julio C. Goez

I deeply appreciate our many rewarding and inspiring conversations over the years, and especially how we worked on problem-solving the autonomous ferry project when the project was at its most difficult. And above all, *thank you* for being an outstanding co-supervisor during my PhD!

Stein Wallace

Very few people have encouragingly challenged my thinking the way Stein does. Ever since I attended his bachelor course in 2015, he has made me reflect in new ways to refine my work or ways of thinking. He was never required to set aside all this time for having these engaging conversations, but he still did.

So thank you for being among the most influential and impactful professors I have ever had, and for being so for nearly a decade. Had it not been for your enthusiastic, clear, and inspiring teaching in that bachelor class in 2015, doing a PhD within operations research would likely never have happened.

Jarle Møen, Kristin Reichel Teigland, Leif Kristoffer Sandal, and Mario Guajardo

The support you have shown me during difficult times in the last years have been invaluable. While some people dislike that I focused on more than just the PhD thesis, you have seen the full picture—that my contributions to society next to the PhD thesis have helped me become a better researcher and ensuring that my work can have a real-life contribution. Thanks to your support, I have been able to realize a greater part of myself. While the last four years have been your impromptu brainstorming session over a lousy hot-pot in Beijing in January 2014[1], and that the decision to study at NHH was made on a whim. So *thank you* for being the friend of a lifetime, for a lifetime.

Kristian Ringen Fauske

No other person has had as much impact on my Bergen years as Kristian. Ranging from ludicrous mountain trips to having had the honor of living with Kristian twice, he has been there through all the lows and all the highs both before and during the PhD. Of course he was there when the decision to apply for a PhD was taken, and of course he was there when I submitted this thesis.

While there are an abundance of examples I could bring up and that I am immensely grateful for, what I would like to highlight is our joint ability in taking an ordinary situation and turning it into something extraordinary and memorable. And as I am about to stalk you down on the fourth continent in less than six years, who knew that just a simple deck of UNO cards could change our lives so dramatically for the better?

Markus N. Torgersen

I got to know Markus much later than Andreas and Kristian. However, his impact in the last years has been so immense that Andreas and Kristian has cheekingly started to look over their shoulders as Markus has catched up with them regarding their impact and importance in my life—the four of us ultimately forming the friend group Mack.

Thank you for an abundance of life-lasting memories. While there's been too many memories I deeply appreciate to be listed here, one particular characteristic of our friendship that I feel particularly grateful for, is our relentless search for understanding the core, underlying principles of whatever rabbit hole of a new, enriching conversational topic that we have entered.

Jørg Sondre Tomstad

Thank you to Jørg not only for our wholesome friendship and support for one another, but also for teaching me an invaluable life lesson I have brought with me since you taught it to me eleven years ago—that it is indeed the devil who is in the details. Moreover, thank you for inspiring me with the valuable lessons of hard work and showing what unconditional inclusion entails. Your

[1] Quite fittingly, this moment was exactly ten years ago!

vast experience has given not only me but also countless of others close to me unforgettable experiences together.

Nikita Dhawan, Vilde Larsen Nakkim, Tora Lindheim, Esra Aslan, Lisa Hellweger, Kristian Ølstad, Peter Jan Svarstad, Jørg Sondre Tomstad, Markus N. Torgersen, Kristian Ringen Fauske, and Andreas Dierkes Landsverk

I would not have had the endurance and zest to undertake a PhD, let alone take on many other challenges in my life during the last decade, had it not been for the greatest group of friends anybody ever had. I have been encouraged, inspired, challenged, listened to, corrected, teased, and tolerated the most by these people during the most defining years of my life. I love you all.

Toni Perry-Thomas

To Toni Perry-Thomas, for months you've shone, your cheer a steady presence as the days have flown. As pages turned and deadlines loomed near, your boundless positivity was constantly there. This acknowledgment, though quiet in its tone, speaks of gratitude for the support you've shown. As I close this chapter, bright and new, know my journey took quite a turn because of you.

The Landsverk Family

I deeply appreciate those who are practically my second family—Andreas, Lærke, Elisabeth, John Petter, Elena, Cecilie, and the rest of the Landsverk family. The way you included me as part of your family almost from the moment I met you has touched my heart tremendously, and I am so excited to share many more memories together! Last, but not least, *thank you* for being there so much for me in the last eight months before I began at NHH—you helped me create a good, healthy foundation I built the most important years of my life on.

Jonas Andersson

I was a blank canvas on everything machine learning until Jonas blasted that door open with such engagement and knowledge in 2018 that it was impossible not to get smitten by it. Thank you so much for going out of your way in creating fantastic content and fostering new data scientists out of NHH, as well as frequently answering my incessant questions on statistics and data science over the years. As data science is the new chapter in my life, the impact of your teaching has already become immense.

Therese Egeland

Thank you for having such a big heart for everyone. Marit Helene Gladhaug and I definitely agree that had the "empathy sponge" prize existed as a complement to the bronze sponge, then you would have won that one as well. Not only are you a tremendous team player that give others opportunities they would not have had without you, but you are an incredible role model for students. You have affected many students' lives even without talking to them because of your knowledge, attitude, attention, energy, and whole-hearted caring in the auditorium. You are a massive inspiration as a teacher and researcher, and I am enormously grateful for our engaging conversations and friendship.

Alexander W. Cappelen

Thank you, Alexander. You inspired me to frequently speak up in Aud Max despite having a 150 BPM heart rate when doing so. For the first time in my life, I witnessed enlightened teaching during your course in 2014. And you showed me that academia can be fun and engaging. Unknowingly to you, you helped me lay a groundwork of skills that has substantially impacted my PhD and non-PhD life since, and I am enormously grateful for that.

Bertil Tungodden, Alexander W. Cappelen, and Erik Sørensen

After working an uncountable number of years as a security guard, Alexander, Bertil, and Erik gave this naïve then-bachelor student the opportunity to have his first relevant job. Despite the fact that I had never written a line of Python, let alone built a web application, all my work would be about building a web application in Python over the course of a couple of months. You threw someone into the deep waters because you believed he would learn to swim quickly. The impact such a decision had on my life is at a very minimum staggering, since technology development and Python and web application skills have been central components of my life ever since—and still is. So *thank you* for giving me the chance—that one hiring decision you made created a whole new trajectory for my life.

Kristian Steffensen

When I was 16 years old, my science teacher Kristian gave us a year-long assignment—to create a binder with all our notes from his course. I was inspired by this challenge, and by the end of the year I entered the teacher's lounge and handed over two volumes and 600 pages of carefully typewritten and digitally illustrated science notes. Kristian was so intrigued by this, he borrowed the binders and read them back to back over the course of a few weeks. The fact that a former Cern researcher could find the incessant effort of a 16-year old to create something worth his attention for days, let alone weeks, was a paradigm shift in my thinking. You, and you alone, gave me the inspiration to try to *build* something at the time. It created a drive and a purpose in life that still drives me. So *thank you*, Kristian—I owe part of this PhD to you.

Victor Norman

While there's a lot to be said about your importance, inspiration, and influence during my bachelor's degree years, there is one situation I want to show particular gratitude for. Thank you for that one lunch during my bachelor's degree in which you listened and then gave me advice for what to do for my master's. Without that one lunch, I would likely never have done a PhD at NHH in the first place.

Ondřej Osička

Ondřej has been my academic big brother who I always looked up to ever since he held lectures in a master's course I was enrolled to. Thank you for inspiring me not only to do a PhD, but also throughout it.

Atle Haugen and Ondřej Osička

One of my proudest non-academic achievements during the PhD years was to complete a 24-hour ultra run. It gave me a rejuvenated feeling of being capable to achieve challenging goals again, and it taught me new principles on overcoming adversity. Principles I have employed when the going got tough during the PhD since. So *thank you*, Atle and Ondřej, not only by inspiring me to join you in the ultra run, but also for being the life of the PhD lunch rooms—you made the PhD life so much easier for all of us.

Eirik Berger

Besides deeply appreciating our frighteningly large overlap of joint interests, our rich conversations, and our great friendship, there is one particular situation I want to show extra appreciation for. During the work on the fourth chapter of this thesis, I ran over two million test instances of an optimization problem. This would have been far from straightforward had it not been for a genius idea from Eirik that saved me weeks or months of frustration.

Roman Kozlov

No other person has had a bigger impact in helping me develop my mathematical skills than Roman. So *thank you* for all the effort you put in to help me develop a mathematical intuition.

Rolf Assev and Sonia Baines

A big thank you to Rolf Assev and Sonia Baines. Without your support and belief in me, I would not have had the chance to explore my love for research-based innovation in the way that I have had.

Åshild Aarø

Thank you, Åshild, for being a caring Leader with a capital L. I am enormously grateful not only for getting to know you during the PhD years, but also for having had the chance to co-found a company with you.

Endre Bjørndal

You are an unsung hero who have multiple times stepped up and helped me since the master's program. Please know how deeply grateful I am for you stepping up and helping out.

Flavio Soriano

A big thank you to Flavio who really stepped up and gave me an invaluable tool kit for my post-PhD career.

André Wattø Sjuve

I deeply appreciate that André made and shared a faultless LaTeX thesis template that saved me numerous hours of work during a hectic time period. I have never experienced as few compilation issues as I have with this template, and that is a testament to the attention and detail you put into creating a template to help others.

Other giants

While this list could easily have been expanded with numerous other people I owe my gratitude to, I would also like to show my deep appreciation for these giants in making a lasting impact in my life—you have all directly or indirectly helped me succeed at the PhD, and I am enormously grateful for each and every one of you. In alphabetical order:

Anders Bratt, Anders Skogseth-Braathen and family, Anders Wettergren Gundersen, Bjarne Braathen and family, Borghild Braathen, Bram Timmermans, The Corellou/Roeleveld family, Coşku Can Orhan, Erlend Dean Unhammer, Eskil Forshaug, Fehime Ceren Ay, Grethe Kristiansen, Hans-Petter Christensen Heine Sunde Nordnes, The Hellweger Family, Henrik Fladvad, Ida Helene Skjeggedal, Ida Matthie, Ivi Hoel Christensen, Jonas Hellandsjø Gjøen, Jørgen Tvedt, Kim Luong, Kjetil Bjorvatn, Kristian Holen, Kristin Ward Heimdal, Lars Christensen, Liv Marit Seides, Mahnaz Fakhrabadi, Maria Aylin Barstad Drageset, Marit Helene Gladhaug, Marius Bratsberg, Martin Boger, Mascha Johanna Fauth, Micaela Andrea Castilla, Ole Kristian Dyskeland, Raman Dhawan, Rasmus Bang, Rezvan Azad Gholami, Samaneh Sheybanivaziri, Sofie Øiestad, Therese Eftevaag, Tor W. Andreassen, Viktoria Eimind, Vilde Koch Fredriksen, Yeojin Kwon, and Ørjan Dale.

Onwards and upwards!

Christian Braathen

Christian Braathen

Bergen, January 2024

Contents

Introduction

Why?

That is the guiding phrase when developing mathematical programming models. Developing a mathematical programming model is not about obtaining the most satisfactory solution among candidate alternatives. Instead, on the one hand, it is about developing insights and understanding *why* the solution is what it is (Geoffrion, 1976). On the other hand, it is about *why* we are interested in the problem in the first place.

A major reason mathematical programming models are worthwhile developing is because they have the potential to improve people's lives directly and indirectly. As a colossal number of people spend substantial parts of their day either working or commuting, even incremental improvements in their work life or commuting experience can significantly impact society.

That is why it is interesting to pursue problems of staffing and transportation. And this thesis is driven by that why.

Most staffing and transportation problems belong under the umbrella term "the assignment problem", which is concerned with matching two or more sets in such a way that an objective is optimized. For two-dimensional problems, these sets are often given the general terms "tasks" and "agents", in which tasks will be assigned to agents. For multi-dimensional problems—problems with more than two sets—frequent additional set types are time slots, vehicles, and alike (Pentico, 2007). While these problems have been thoroughly studied and developed since the seminal work of Kuhn (1955), advances in algorithmic efficiency, computational power, technology in general, legislation, and new applications—requiring extensions of existing models—continuously open new, exciting research questions.

The thesis is structured in four chapters, each an independent article. The first three articles of this thesis are directly attributable to the assignment problem, while the fourth is concerned with a related data availability issue. The remainder of this introduction describes these chapters in detail in three stages. First, each chapter's research question is presented. Then, Table 1 and

2 describe a high-level overview of each chapter's methodology, focus areas, and classification codes. Finally, each chapter is summarized.

1. Chapter 1 is a resource allocation problem and answers, "How to schedule mass interviews in a volunteer-driven organization, ensuring favorable creation of individual and paired interviewer schedules?"

2. Chapter 2 is a logistics problem and answers, "How do social constraints—and overtime considerations in particular—affect the selected routes and the quality of the solution?"

3. Chapter 3 is a public transportation problem and answers, "What are the operational implications and benefits for passengers when ferries are operated autonomously instead of manually?"

4. Chapter 4 is concerned with suppressed data points in commuting data and answers, "How to reconstruct the trip-distribution matrix from commuting trip data with cell suppression, and what are the impacts on estimates and predictions due to such data limitations?"

Table 1. Methodology and focus areas.

	Chapter 1	**Chapter 2**	**Chapter 3**	**Chapter 4**
Aspect	Interview Scheduling	Vehicle Routing	Autonomous Ferries	Commuting
Optimization focus	Multi-dimensional assignment	Multi-dimensional assignment	Multi-dimensional assignment	Prediction
Optimization method	Integer linear programming	Mixed integer linear programming	Integer linear programming	Nonlinear and mixed integer linear programming
Solving approach	Exact	Exact and heuristic	Exact and heuristic	Exact
Application domain	Resource allocation	Logistics and distribution	Public transportation	Commuting
Staffing focus	High	High	Medium	None
Transportation focus	None	High	High	High
Number of variables	Medium–high	High	High	Low
Number of constraints	Medium–high	Low	High	Low
Data Source	Real-world	Real-world	Synthetic	Synthetic
Case Study Location	Norway	Norway	Norway	Sweden
Target Audience	Volunteer organizations	Logistics companies	Public transport authorities	Statistical agencies
Key Findings	Scheduling efficiency	Overtime usage and cost reductions	Vessel utilization	Privacy issues

Table 2. This thesis centers primarily around the JEL classification codes C44, C61, J22, and R41.

JEL code		**Chapter 1**	**Chapter 2**	**Chapter 3**	**Chapter 4**
C44	Operations Research	✓	✓	✓	✓
C61	Optimization Techniques	✓	✓	✓	✓
J22	Time Allocation and Labor Supply	✓	✓	✓	
R41	Transportation Demand		✓	✓	✓
M54	Labor Management	✓			
L81	Retail and Wholesale Trade		✓		
R23	Regional Labor Markets				✓
C15	Statistical Simulation Methods: General				✓
C81	Methodology for Collecting, Estimating, and Organizing Microeconomic Data				✓
C82	Methodology for Collecting, Estimating, and Organizing Macroeconomic Data				✓

Chapter 1. Interview Scheduling Using Integer Linear Programming

Volunteerism is a substantial cogwheel in modern-day societies, with an estimated one billion volunteers worldwide (United Nations Volunteers programme, 2018). Volunteers achieve direct career benefits from volunteering, with volunteers having a 27% higher chance to be employed than people who do not volunteer (Corporation for National and Community Service, 2013). This number may partly be explained by the fact that 92% of human resource executives consider volunteering to improve leadership skills (Deloitte, 2016).

At the Norwegian School of Economics (NHH), the student association organizes three projects that hire approximately 3000 volunteer students through three hiring rounds each over the projects' two-year duration. Each project consists of numerous groups in which each volunteer applies to two. The student association requires that every student applying to volunteer in these projects be interviewed, and the interviewers must represent the two groups the student applied to volunteer within.

This requirement consequently makes the scheduling of these hiring rounds difficult, and sub-optimal interviewing schedules then cause poor time utilization during the recruitment process for the interviewers. Poor time utilization is a frequently cited reason why students refrain from participating in such projects. One organizer stated, "In my experience, it is *especially* the recruiting part of student association work most people emphasize when explaining why they do not have the time to engage in [the student association]".

Moreover, since the volunteers do not receive a salary for their efforts (Shin & Kleiner, 2003), it is imperative to delegate tasks and responsibilities they prefer to remain motivated (Gordon and Erkut, 2004; Sampson, 2006). Consequently, it is valuable to provide these project organizations with better interviewing schedules to increase the potential of volunteerism in the student association.

This article aims to contribute in three primary ways. First, the article contributes to the practice of operations research by applying the model to the scheduling challenges faced by NHH's student association, in addition to reporting and discussing the results of this application. Second, and in alignment with the first aim, the applied model attempts to increase awareness of, and enthusiasm about, operations research in general to the student mass since the model is employed in a context the majority gathers around—the student association. Hence, this article focuses on applying and implementing practical operational research problems. Third, the article provides new features to the meeting scheduling problem that have yet to be used in the literature.

This work has been presented at "the 22nd Conference of the International Federation of Operational Research Societies (INFORMS)."

Chapter 2. Assessing the Impact of Driver Overtime in the Distribution Network of a Flower Retail Chain

with Mario Guajardo

The Vehicle Routing Problem (VRP) represents an essential challenge in transporting goods and services, with its fundamental objective being to identify an optimal set of routes that service a set of customer demands (Adewumi & Adeleke, 2018).

The most common objective of real-world VRP variants is to minimize total distribution costs while upholding high-quality distribution services (Konstantakopoulos et al., 2020). While driver wages are a substantial component of total distribution costs, drivers receive little focus in VRP studies. Specifically, in the taxonomy of Tan and Yeh (2021), only two components explicitly mention drivers, and only a minority of VRP articles include working hour considerations and legislation protecting the drivers—such as required breaks when driving throughout the day (Lahyani et al., 2015).

Consequently, while the vehicle routing problem (VRP) has been thoroughly studied, a large share of studies ignore social constraints (Lahyani et al., 2015). These constraints relate to the drivers and are important for several reasons. First, driver safety is a real-life concern and should be considered for practical considerations in the VRP. Second, implementing social constraints impacts total transportation costs, total driving distance, and other commonly used objective function values due to their constraining nature. Third, social considerations may be substantially impacted by each type of objective function unless constrained sufficiently. Hence, we need to understand to which extent these considerations are impacted.

Although the concepts of social constraints are acknowledged by the practitioners in charge of routing, it can be challenging to grasp all the details of the routes, the trade-offs between them, and the implications of social constraints for the total costs. In practice, planners tend to manually tweak the software solutions to cope with these social constraints or manually do the whole routes from scratch. With access to real data instances and both the practitioners' and commercial software's solutions, we have studied in detail the impact of social constraints in this VRP, with a particular focus on overtime usage.

This article aims to contribute in three primary ways. First, the article demonstrates how social constraints, and overtime usage in particular, are affected by commonly used objective functions. Second, the article provides an intuitive routing generation technique that is explainable to experts. Third, the article provides clear recommendations to practitioners to employ a more flexible policy for overtime usage—but within the allowable working day duration.

This work has been presented at "the 17th CEMS Research Seminar on Supply Chain Management" and is accepted for presentation at "the 9th International Workshop on Freight Transportation and Logistics (Odysseus 2024)."

Chapter 3. Autonomous Ferries in Light of Labor Regulations—A Passenger Perspective

with Julio C. Goez and Mario Guajardo

Similar to many urban areas, Bergen faces transportation challenges. With a growing population, a surge in tourism, and public transportation that resembles a hub-and-spoke structure—with most buses going via the city center due to the coastline and mountainous surroundings—the city's transportation system is congested and under pressure to accommodate the increasing demand for mobility. This challenge spurred the interest of the "blue light rail" project. The project aims to use autonomous ferries to provide an alternative mode of transportation for the city's residents and tourists to reduce traffic congestion and improve the overall transportation experience.

The use of autonomous vessels (AVs) in general—some of which are ferries—is a highly topical issue due to the increasing capabilities of artificial intelligence and changing regulations allowing their use in national and international waters. While autonomous ferries have their appeal, an alternative is using manually operated ferries to reduce capital expenditures. Hence, a highly topical question is to understand how much better autonomous ferries can meet passenger demand than manually operated ferries subjected to labor regulations.

In this article, we demonstrate that using autonomous ferries in public transportation can significantly improve the passengers' utility, particularly in Norway, where there is a high coast-to-land ratio, road infrastructure is expensive, and a positive outlook on AVs by authorities makes this technology particularly relevant. Due to potential practical limitations in the number of operating hours for autonomous ferries and due to other transportation alternatives that will operate the most during high-demand hours, our results indicate that the best usage of autonomous ferries is during low-demand hours of the day and week—in other words, from late evenings to early mornings in addition to weekends.

This article aims to contribute to the literature in two primary ways. First, to understand to which extent autonomous ferries may outperform manually operated ferries regarding passenger benefits. And second, to suggest how autonomous ferries should be employed in a system of other available transportation alternatives.

This work has been presented at "the NORS Annual Conference 2021" and "the 10th INFORMS Transportation Science and Logistics Society Workshop."

Chapter 4. Adjusting for Cell Suppression in Commuting Trip Data

with Inge Thorsen and Jan Ubøe

The increasing access to micro-data has reinforced issues related to privacy and confidentiality. As pointed out by Matthews and Harel (2011), data-releasing agencies must account for the fact that privacy is viewed as a fundamental human right by the United Nations. This perspective explains why statistical agencies often introduce limitations in releasing data, of which cell suppression is the most commonly used approach for tabular data (Abowd & Schmutte, 2016).

For example, Statistics Norway demands that at least three units have to underlie a total value to be published in a cell to avoid the risk of a residual disclosure.

However, suppressing cells in commuting data provides problems. Particularly for sparsely populated areas, the cell suppression may occur in a considerable number of cells and potentially lead to seriously biased estimates—for example, on how variations in distance affect the travel demand.

Additionally, more than suppressing data in cells is needed to protect privacy rights. Specifically, our approach succeeds to a large degree in disclosing the information hidden by suppressing information in cells with less than three commuters. This information can then be linked to information in other databases that can then be used to identify individuals and more sensitive attributes. Consequently, suppressing cells in commuting data not only leads to biased estimates but also does not achieve their intended privacy protection intentions.

This article aims to contribute in two primary ways. First, we demonstrate that suppressing the information in cells with a low number of observations does not necessarily preserve privacy adequately. The agencies should consider suppressing marginal totals, using higher cut-off values, and/or other methods, such as noise infusion. Second, our results strongly recommend that researchers develop sound methods to adjust for suppressed information rather than just ignoring the cells by setting the values equal to zero—by doing so, we largely removed the estimation bias of the distance deterrence parameter in a doubly constrained gravity model.

References

Abowd, J. M., & Schmutte, I. M. (2016). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, *2015*(1), 221–293.

Adewumi, A. O., & Adeleke, O. J. (2018). A survey of recent advances in vehicle routing problems. *International Journal of System Assurance Engineering and Management*, *9*, 155–172.

Corporation for National and Community Service. (2013, June). Volunteering as a pathway to employment: Does volunteering increase odds of finding a job for the out of work? https://americorps.gov/sites/default/files/%5C%5Cevidenceexchange/FR_2013_VolunteeringasaPathwaytoEmployment_1.pdf

Deloitte. (2016, June). Building leadership skills through volunteerism.

Geoffrion, A. M. (1976). The purpose of mathematical programming is insight, not numbers. *Interfaces*, *7*(1), 81–92.

Gordon, L., & Erkut, E. (2004). Improving volunteer scheduling for the Edmonton Folk Festival. *Interfaces*, *34*(5), 367–376.

Konstantakopoulos, G. D., Gayialis, S. P., & Kechagias, E. P. (2020). Vehicle routing problem and related algorithms for logistics distribution: A literature review and classification. *Operational research*, 1–30.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83–97.

Lahyani, R., Khemakhem, M., & Semet, F. (2015). Rich vehicle routing problems: From a taxonomy to a definition. *European Journal of Operational Research*, *241*(1), 1–14.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1–29.

Pentico, D. W. (2007). Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, *176*(2), 774–793.

Sampson, S. E. (2006). Optimization of volunteer labor assignments. *Journal of Operations Management*, *24*(4), 363–377.

Shin, S., & Kleiner, B. H. (2003). How to manage unpaid volunteers in organisations. *Management Research News.*

Tan, S.-Y., & Yeh, W.-C. (2021). The vehicle routing problem: State-of-the-art classification and review. *Applied Sciences*, *11*(21), 10295.

United Nations Volunteers programme. (2018). The thread that binds–volunteerism and community resilience.

Interview Scheduling Using Integer Linear Programming

Christian Braathen

Abstract

This article applies the meeting scheduling model to schedule en-masse interviews over multi-day events. Creating such meeting schedules is a complicated, time-consuming task that usually requires considerable manual planning when hundreds of meetings are involved and if the quality of the schedule is a priority. The problem consists of three overlapping and gradually more difficult objectives—that is, at favorable times, to assign meetings, assign individual interviewers consecutive meetings, and assign interviewer pairs consecutive meetings. The model has been applied five times within a student organization, ranging from 17 to 1149 meetings to be scheduled. Compared to former, manually created schedules, both the organization's and the interviewers' perceived interviewing experience is markedly improved.

Keywords: meeting scheduling; assignment problem; integer linear programming; volunteer management; OR practice

JEL Classification: **J22** Time Allocation and Labor Supply; **M54** Labor Management; **C44** Operations Research; **C61** Optimization Techniques

1.1 Introduction

Volunteerism is a substantial cogwheel in modern-day societies, with an estimated one billion volunteers worldwide (United Nations Volunteers programme, 2018). The United Nations describes it as the thread that binds communities together (United Nations Volunteers programme, 2018), and volunteers achieve direct career benefits from volunteering. Specifically, volunteers have a 27% higher chance to be employed than people who do not volunteer (Corporation for National and Community Service, 2013), which may in part be explained by the fact that 92% of human resource executives consider volunteering to improve leadership skills (Deloitte, 2016). However, since the volunteers do not receive a salary for their efforts (Shin & Kleiner, 2003), it is imperative to delegate tasks and responsibilities they prefer to remain motivated (Gordon and Erkut, 2004; Sampson, 2006).

One organization focusing extensively on providing tasks and responsibilities the volunteers prefer is among Norway's most active student associations. At the NHH Norwegian School of Economics (NHH), students operate an association consisting of 120 subgroups as well as three larger projects every two years. These projects hire approximately 3000 volunteer students through three hiring rounds each over two years—arranging the second largest cultural festival in Western Norway, Northern Europe's largest student-driven business conference, and a five-day student sporting event for about 2000 participants. Since the projects are time-limited and volunteered by students wanting to volunteer in different positions with different responsibilities throughout their studies, the volunteer turnover rate is practically 100 percent. Hence, the student association must recruit about 3000 volunteer students anew every two years during a few multi-day events recruiting and interviewing the students.

To be inclusive, the student association requires that the project organizations interview every student applying to become a volunteer and that managers for the applicants' preferred positions must conduct the interviews. This concern complicates the scheduling process as only a selected few people can conduct each interview. However, the need to work with preferred tasks and responsibilities extends beyond applicants. Specifically, the interviewers are also volunteers, and their past unsatisfactory interviewing schedules made them less willing to take on such volunteer positions. Hence, the student association has expressed a wish for a decision support tool that can provide better schedules for the interviewers while respecting the preferences of the applicants.

This article aims to contribute in three primary ways. First, the article contributes to the practice of operations research by applying the model to the scheduling challenges faced by NHH's student association, in addition to reporting and discussing the results of this application. Second, and in alignment with the first aim, the applied model attempts to increase awareness of, and enthusiasm about, operations research in general to the student mass since the model is employed in a context the majority gathers around—the student association. Hence, this article focuses on applying and implementing practical, operational research problems. And third, the article provides new features to the meeting scheduling problem that has not been used in the literature yet.

The article is organized as follows. Section 1.2 explores the interview allocation problem together with the meeting scheduling problem. Section 1.3 presents a case study while the corresponding optimization model and results are presented in Section 1.4 and 1.5, respectively. The article concludes in Section 1.6 and highlights how the meeting scheduling problem has become more relevant in the post-COVID-19 era and that this increased relevance can be utilized to improve meeting scheduling within organizations.

1.2 Literature

The interview scheduling problem is a subset of the closely related meeting scheduling problem between two or more parties, except that a secretariat—consisting of homogeneous interviewers who can attend any meeting—may participate in interviews if necessary. Due to the similarity between the problems, the literature on the meeting scheduling problem should be studied, too. Moreover, both problems can be considered part of a larger class of problems—namely, the assignment problem. Such problems involve assigning parties and potentially assigning parties to a specific time. The parties in assignment problems can be animate objects (agents), inanimate objects (tasks), or both.

In comparison, the meeting scheduling problem involves that individual $i = 1, ..., n$ should participate in (up to) a pre-defined, varying number of meetings m_i allocated at certain times during t time slots, being restricted by having up to r simultaneous meetings due to location constraints—such as the number of rooms. While there is usually a substantial number of meetings to be assigned—up to $\sum_{i=1}^{n} m_i$—there are usually far fewer meetings than possible meeting starts. This flexibility, coupled with the fact that there are usually only a few parties involved in each meeting, creates a focus on the individuals, allocating their respective meetings at favorable times without wasting their time on unnecessary waiting.

Moreover, while meeting scheduling problem resembles personnel scheduling, the problems are still distinctly different. On the one hand, for both problems, there are people to be scheduled for specific time blocks, often involving particular preferences, and there exist overarching organizational goals and multiple stakeholders with individual priorities. On the other hand, personnel scheduling is more focused on economic considerations—such as minimizing wages—and respecting contractual and legislative constraints. Additionally, planners of personnel schedules are often involving overlapping shifts to meet fluctuating demand. Ultimately, these concerns increase the scheduling complexity compared to meeting scheduling problems. However, meeting scheduling is more concerned about scheduling specific people to be at the same time and place together than personnel scheduling. Such coupling considerations consequently increase the scheduling complexity substantially compared to personnel scheduling. Hence, these two problems are distinctively different despite their overlapping similarities. For a thorough review of personnel scheduling literature, see Van den Bergh et al. (2013).

A broad overview of the meeting scheduling literature and key model features are presented in Table 1.1, and selected key features are described next.

Table 1.1. Meeting scheduling literature comparison. Model features highlighted in bold are covered only by the model presented in this article.

Model feature	P	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
More than two parties can meet	C	N	N	C	N	N	N	N	N	N	C	N	N
>1 representatives inside each party	C	N	C	C	N	C	N	N	C	N	N	N	C
Various time slots per meeting	C	N	N	N	N	N	N	N	N	C	N	N	C
Multiple days constraints	C	C	N	N	N	N	N	N	N	C	N	N	N
Infrastructural constraints	C	N	N	C	I	C	C	N	C	I	I	N	N
Consecutive meetings	C	I	I	C	I	I	I	I	I	N	N	N	N
Meeting preferences	C	N	N	N	N	C	C	C	C	N	N	C	C
Multiple venues	C	N	N	N	N	N	N	N	N	N	N	N	N
Schedule all meetings	C	N	C	C	C	C	N	N	N	C	C	C	N
Preserving parties	C	N	N	N	N	N	N	N	N	N	N	N	N
Time preference	C	N	N	N	N	N	C	N	C	N	N	C	N
Unavailability considerations	C	N	N	C	C	C	N	N	C	N	N	C	C
Enforced breaks	C	N	N	N	N	N	C	N	N	N	N	C	C
Seniority	C	N	N	N	N	N	N	N	N	N	N	N	N
Reduce room changes	I	N	N	N	N	C	C	N	N	N	C	N	N
Minimize resource use	I	N	N	N	N	C	C	N	N	N	N	N	N
Moving constraint	N	N	N	N	N	C	C	N	C	N	N	N	N
Fair allocation	N	N	N	N	N	N	N	N	N	N	N	N	N

C: model feature considered; I: model feature implicitly considered; N: model feature not considered.

P: the model presented in this article; [1]: Bartholdi III and McCroan (1990); [2]: Rinaldi and Serafini (2006); [3]: Pesant et al. (2015); [4]: Le Roux et al. (2015); [5]: Gebser et al. (2013); [6]: Gueret et al. (2009); [7]: Huang et al. (2012); [8]: Ernst et al. (2003); [9]: Mausser et al. (1996); [10]: Kiyonari et al. (2006); [11]: Aizam and Sim (2016); [12]: Schrage (2004).

Since meeting scheduling primarily focuses on the individuals, improving their well-being is essential. Four factors are essential to provide a foundation for the participants' well-being—namely, to consider the preferences of which parties should meet, to respect both time preferences and unavailability considerations, and to prevent the meeting participants from having meetings at different venues over a short time-frame.

Thus far, the literature has inconsistently covered these four foundational well-being considerations. Specifically, considering the preferences of which parties should meet is covered by Gebser et al. (2013), Gueret et al. (2009), Huang et al. (2012), and Ernst et al. (2003), while only Gueret et al. (2009) and Ernst et al. (2003) provide time preferences. Moreover, unavailability considerations are considered by Pesant et al. (2015), Le Roux et al. (2015), Gebser et al. (2013), and Ernst et al. (2003), while the existing literature does not cover multiple venues yet.

Moreover, additional considerations should preferably be made to further improve the participants' well-being. One such consideration is to maximize the consecutive interviews of the interviewers, which has been the most frequent objective in the meeting scheduling literature.

Despite the importance of this consideration, it has usually been modeled with a proxy objective. For instance, Bartholdi III and McCroan (1990) describe the need for consecutive interviews but attempt to model it implicitly by allocating meetings sooner. Rinaldi and Serafini (2006), Gebser et al. (2013), and Le Roux et al. (2015), in comparison, achieved consecutive interviews implicitly by reducing the number of idle time slots.

However, it may be better to reduce the number of breaks instead of time slots. For instance, the meeting participants may consider a long break of three time slots as more favorable than three breaks of one time slot each—especially if the time slot is short. The project organizations in this article favor the approach of reducing the number of breaks, which was also the objective of Pesant et al. (2015).

Reducing the number of breaks brings another key benefit, too. Specifically, when multiple parties attend multiple meetings together, which was also considered by Pesant et al. (2015), then the participants' well-being can be improved further if they can undertake multiple meetings consecutively together. This effect can be described as "preserving" the meeting parties, and such preservation has not been covered by the existing literature yet. However, this lack of coverage occurs primarily because most of the literature schedules meetings of two parties only—rendering the preserving-participants consideration irrelevant.

Furthermore, while a particular focus on the individuals is important in a meeting scheduling model, addressing differences in meeting properties is another important concern. Specifically, such differences is usually achieved both by allowing a varying number of parties to attend each meeting and by planning with a varying meeting duration. For the former matter, most of the literature studies cases in which only two parties will attend each meeting (Bartholdi III and McCroan, 1990; Rinaldi and Serafini, 2006; Le Roux et al., 2015; Gebser et al., 2013; Gueret et al., 2009; Huang et al., 2012; Ernst et al., 2003), while the latter has largely been ignored by the literature with the exception of Mausser et al. (1996) and Schrage (2004).

1.3 Background

At NHH, the student association organizes three projects that hire approximately 3000 volunteer students through three hiring rounds each over the projects' two-year duration. Each of these projects consists of numerous groups in which each volunteer applies to two. The student association requires that every student applying to become a volunteer in these projects is interviewed, and the interviewers must represent the two groups the student applied to volunteer within. If a student applies to groups led by the same set of managers, then the venue's secretariat—whose homogeneous interviewers can attend any meeting if needed—will take on the role of the second interviewer.

Interviews should preferably be scheduled consecutively and early in the week. Specifically, the association prefers that interviewers undertake as many of their interviews as possible early on in the interviewing schedule so that the student association has more flexibility in reallocating interviews if necessary. Within a given day, it is generally favorable to undertake interviews early. The only exception to this intraday preference is that the earliest time slots of 8–9 AM are less favored than 9–10 AM but equally favored as 10 AM–4 PM—see Table 1.2 for more details. Moreover, the interviewers should preferably conduct interviews consecutively because the interviewers can then utilize their time more efficiently, having more time to study without frequent interruptions to conduct interviews.

Table 1.2. Time preference ranks, exemplified with a three-day meeting window.

		Day		
From	To	1	2	3
08:00 AM	09:00 AM	2	4	5
09:00 AM	10:00 AM	1	2	3
10:00 AM	04:00 PM	2	4	5
04:00 PM	06:00 PM	3	5	7
06:00 PM	08:00 PM	4	6	

These interruptions are a common concern among the students, and poor time utilization during the recruitment process is an oft-cited reason why students do not participate. One organizer stated, "In my experience, it is *especially* the recruiting part of student association work most people emphasize when explaining why they do not have the time to engage in [the student association]". These considerations, coupled with the fact that the students are allocated to responsibilities they prefer and the leaders of these responsibilities must conduct these interviews, cause substantial overhead efforts that must be made to plan these interviews.

However, early and consecutive meetings are merely a prerequisite for the student association. As described in Section 1.2, certain foundational factors are essential to include to ensure a satisfying schedule. For the student association, having two parties that will conduct multiple interviews with a different set of applicants of various duration is a factor that makes the scheduling process particularly complicated. Additionally, the planners should respect interviewers' unavailability, ensure that the more senior group member undertakes the interview if (s)he is available because each group may bring both a group leader and deputy leader(s) to undertake interviews, and that no interviewers should need to travel to multiple venues within the same day to undertake interviews. Moreover, all applicants must be interviewed to avoid making any students feel excluded from the otherwise inclusive student association. Since the student association must reserve rooms to hold interviews, there may be a varying amount of rooms available at any one time, which creates an upper limit on the number of simultaneous interviews that can be conducted.

Besides allocating more early and consecutive meetings by applying the optimization model presented in this article, the student association also wishes for other qualitative improvements. Specifically, they would like to ensure that interviewers get some resting breaks between consecutive interviews so that they do not, for instance, need to interview for eight hours straight on the first day of the scheduling period. If an interview cannot be scheduled consecutively, the association prefers that the interview is still scheduled for a favorable time. However, for interviews that *can* be scheduled consecutively, the student association appreciates if the interviewer pairs are preserved across consecutive interviews because they already share a meeting room. However, preserving pairs have a lower priority to the student association than achieving consecutive interviews.

The project organizations had the following workflow. First, they provided the input data in a spreadsheet, specifying information about the interviewers, applicants, what constitutes meetings, room availability, and time periods to conduct interviews. A meeting in this context is a pre-defined decision about which two interviewers will meet each applicant and for how long. Additionally, the organizers specified when the interviewers were unavailable in the spreadsheet. The information was then converted into sets and parameters and subsequently added to the optimization model described in the next section. Afterward, the produced schedules were converted to spreadsheets—both individual per interviewer and collective—for distribution and administration.

Providing better schedules have beneficial implications for the association and its students. Specifically, an optimization model that can help students with this overhead task will give other students better working conditions when undertaking interviews and may also contribute to more students wanting to volunteer and take on managerial roles within the association. Since many of these students will take on future leading positions in the business landscape, and with executives believing leadership skills improve with volunteering (Deloitte, 2016), the benefits of such a decision tool extend beyond the direct effects created within the volunteer organizations.

1.4 Integer Linear Programming Formulation

The constraints and objective function presented in this section consist of the sets, subsets, parameters, and variables that are listed in Table 1.3.

Table 1.3. Model components.

Sets	
M	Set of meetings.
T	An ordered set of time slots.
P	Set of participants.
D	Set of days.
L	Set of venues.
G	Set of groups.
$\mathbb{I}^{\mathcal{C}}$	Set of $\mathcal{C}$'s indices—$\mathbb{I}^{\mathcal{C}} := \{i + 1 \in \mathbb{N}_{\lvert\mathcal{C}\rvert}\}$.
Subsets	
$\mathcal{B}_m \subseteq T$	An indexed set of time slots an interview $m \in M$ may begin.
$\mathcal{I} \subseteq P$	The set of interviewers.
$\mathcal{A} \subseteq P$	The set of applicants.
$\mathcal{S} \subseteq P$	The set of the venue secretariat.
$\mathcal{P}_p \subseteq M$	An indexed set of meetings that participant $p \in P$ may attend.
$\mathcal{T}_d \subseteq T$	An indexed set representing which time slots occur in day $d \in D$.
$\mathcal{L}_l \subseteq M$	An indexed set representing which meetings take place on venue $l \in L$.
$\mathcal{F} \subseteq T$	Set of final time slots each day.
$\mathcal{C}_c \subseteq T$	Indexed set of all intraday consecutive time slots from time c to time $c + b$.
$\mathcal{P} \subseteq \{\mathcal{I}, \mathcal{I}\}$	The tuple set of interviewer pairs (i, j) who can conduct interviews together.
$\mathcal{G}_g \subseteq \mathcal{I}$	The ordered set of interviewers that belong to group $g \in G$.
$\mathcal{D}_t \subseteq D$	An ordered subset that lists the day that time slot $t \in T$ belongs to.
Parameters	
$d_m \geq 0$	The number of time slots needed for an interview $m \in M$.
$r_{tl} \geq 0$	Maximum number of rooms available at time $t \in T$ on venue $l \in L$.
$b \geq 0$	Maximum number of time slots an interviewer may undertake consecutively.
$s_i \geq 0$	How many simultaneous interviews that interviewer $i \in \mathcal{I}$ can conduct.
$u_{it} \in \{0, 1\}$	A binary parameter denoting whether interviewer $i \in \mathcal{I}$ is unavailable at time $t \in \mathcal{T}$.
$p_t \geq 0$	Preference for undertaking an interview at time t.
$\omega^A, \omega^C, \omega^P \geq 0$	Weights for the assignment, consecutiveness, and preservation objectives, respectively.
Variables	
$x_{mt} \in \{0, 1\}$	Meeting $m \in M$ is ongoing at time $t \in T$.
$y_{mt} \in \{0, 1\}$	Meeting $m \in M$ begins at time $t \in T$.
$z_{ild} \in \{0, 1\}$	Interviewer $i \in \mathcal{I}$ undertakes meetings at venue $l \in L$ on day $d \in D$.
$v_{it} \in \{0, 1\}$	Interviewer $i \in \mathcal{I}$ attends two consecutive meetings at time $t \in T$ and $t + 1 \in T$.
$w_{ijt} \in \{0, 1\}$	Interviewer pair $(i, j) \in \mathcal{P}$ is preserved from time $t \in T$ to $t + 1 \in T$.

1.4.1 Time Constraints

We must ensure that the time-related constraints are respected. First, we define that each applicant should have only one interview, equivalent to having one interview start. Additionally, all slots in a multi-slot interview must occur in one day instead of spanning over two days.

Hence:

$$\sum_{m \in \mathcal{P}_a} \sum_{t \in \mathcal{B}_m} y_{mt} = 1 \quad \forall a \in \mathcal{A} \tag{1.1}$$

Second, we must define when the meeting starts and does not start, respectively, to help define that the d_m time slots are assigned consecutively:

$$\sum_{t=\tau}^{\tau+d_m-1} x_{mt} \leq (d_m - 1) + y_{m\tau} \quad \forall m \in M, \tau \in \mathcal{B}_m \tag{1.2}$$

$$\sum_{t=\tau}^{\tau+d_m-1} x_{mt} \geq d_m y_{m\tau} \quad \forall m \in M, \tau \in \mathcal{B}_m \tag{1.3}$$

Third, we define the meeting's duration as the number of time slots to be allocated to a meeting across all time slots. This constraint is necessary as (1.2) and (1.3) are summed only over a subset of time slots.

$$\sum_{t \in T} x_{mt} = d_m \quad \forall m \in M \tag{1.4}$$

Handling Infeasibilities

Test instances may be infeasible primarily for two reasons. First, the number of people applying to meet a selected interviewer may be greater than the time slots available. Second, the interviewers may have declared too many time slots in which they are unavailable that no feasible solution exists. The latter reason becomes more prevalent both the more unavailable a meeting participant is and the more people that attend meetings if they have all declared some unavailability. Additionally, it is generally harder to find a feasible schedule if time unavailability spans whole days—for example, two interviewers who are unavailable in the first and second half of the scheduling period render a schedule infeasible if they are supposed to attend at least one meeting together.

In order to handle infeasibility issues if they occur, constraints (1.1) and (1.4) must be modified.

Specifically, constraint (1.1) must be rewritten to a smaller-than-or-equal-to constraint to allow some meetings to not start at all. Additionally, we must rewrite constraint (1.4) to specify that we either assign all or none of a meeting's time slots, depending on whether a meeting has been defined to start or not, respectively:

$$\sum_{t \in T} x_{mt} = d_m \sum_{\tau \in \mathcal{B}_m} y_{m\tau} \quad \forall m \in M \tag{1.5}$$

The choice of handling infeasibilities on an as-needed basis instead of by default is deliberate. Specifically, if fewer than all interviews could be scheduled from the onset, an optimal solution could allocate fewer than all interviews to schedule more consecutive interviews or preserve interviewer pairs across consecutive interviews. Since the primary aim of the project organizations is to schedule every applicant for an interview, it is preferred to prioritize scheduling everyone first and then focus on favorable allocations.

It could be possible to either penalize or reward the objective function for missing allocations or for allocating everyone, respectively. However, infeasibility is not a big concern for most test instances presented in this article because, after the first test instance presented in Section 1.5, the project organizations were told to allow only limited unavailability for each interviewer to ensure that the primary aim—to schedule all interviews—could be achieved. Hence, for the test instances presented in this article, it is wise to handle infeasibilities if they occur instead of handling such concerns to the model outright.

1.4.2 Infrastructural Constraints

Interviewers may conduct interviews at multiple venues. Due to the geographical distances between the venues, each meeting is pre-defined at a specific location to provide the applicants with short traveling distances. For such cases, the project organizations specify that they want every interviewer to undertake interviews at maximally one venue only per day, while the venue that an applicant will be interviewed at is pre-specified in the application form.

We must hence define if an interviewer i is at venue l on day d:

$$\sum_{m \in \mathcal{P}_i \cap \mathcal{L}_l} \sum_{t \in \mathcal{T}_d} x_{mt} \leq |\mathcal{T}_d| z_{ild} \quad \forall d \in D, l \in L, i \in \mathcal{I} \tag{1.6}$$

If meetings had been undertaken with only two parties, then the allocation problem could have been split along the location dimension into multiple, smaller problems, potentially simplifying the solving process. However, since an arbitrary number of participants can meet and since the participants may meet on multiple occasions and locations, we may usually not be able to split the problem along the location dimension. Therefore, we specify that an interviewer can maximally be at one venue on any given day:

$$\sum_{l \in L} z_{ild} \leq 1 \quad \forall d \in D, i \in \mathcal{I} \tag{1.7}$$

Next, as the meetings usually occur physically, there is a constraint on the number of rooms available at time t and venue l:

$$\sum_{m \in \mathcal{L}_l} x_{mt} \leq r_{tl} \quad \forall t \in T, l \in L \tag{1.8}$$

1.4.3 Consecutive Interviews Constraints

A key request from the project organizations was to ensure that interviewers could undertake consecutive interviews. Then the constraint specifying when consecutive time slots with allocated interviews are *not* achieved is defined as:

$$\sum_{m \in \mathcal{P}_i} x_{mt} + \sum_{n \in \mathcal{P}_i} x_{n,t+1} - \sum_{o \in \mathcal{P}_i} \sum_{\substack{\tau = \max(1, t - d_o + 2): \\ d_o > 1}}^{t} y_{o\tau} \geq 2v_{it} \quad \forall i \in \mathcal{I} \setminus \mathcal{S}, t \in T \setminus \mathcal{F} : |\mathcal{P}_i| > 1 \tag{1.9}$$

The left-hand side of (1.9) studies whether an interviewer i participates in two meetings at time t and $t+1$. However, a meeting that lasts longer than one time slot may have started at time t or before and still be ongoing at time $t+1$. To avoid defining such a meeting as a consecutive meeting, we need a correction term. The correction is represented by the third term on the left-hand side in (1.9), which will sum to one if a multi-slot meeting o that interviewer i is participating in is ongoing at time t but will end strictly after time t, and sum to zero otherwise. Additionally, since only consecutive time slots within a day are interesting, we apply this constraint only for intraday consecutive slots, excluding end-of-day time slots $t \in \mathcal{F}$.

It is worth noting, however, that the constraint applies only to non-secretariat interviewers because the secretariat consists of multiple members that must be present either in a meeting or in the secretariat's office that is located near the meeting rooms. Hence, there is no need for consecutive interviews for the secretariat, and so the constraint applies for $i \in \mathcal{I} \backslash \mathcal{S}$.

Interviewing breaks

Another concern from the project organizations was that if the optimization model performed well enough to assign a multitude of consecutive interviews, the interviewers must also get occasional breaks of at least one time slot. This concern does not apply to the secretariat as they consist of multiple people.

$$\sum_{m \in \mathcal{P}_i} \sum_{t \in \mathcal{C}_c} x_{mt} \leq b \quad \forall i \in \mathcal{I} \backslash \mathcal{S}, c \in \mathbb{I}^{\mathcal{C}} \tag{1.10}$$

1.4.4 Preserving Interviewer Pair Constraints

After a dry-run producing a schedule, a project organization specified that it would be beneficial if interview pairs could be preserved from one consecutive interview to another. Preserving the interviewer pairs consequently made it easier for the interviewer pairs to compare candidates as candidates more frequently were scheduled consecutively.

Similarly to (1.9), a correction term is needed to exclude each multi-slot interview that started at time t or before and continues at time $t+1$. We can then define a constraint specifying that an interviewer pair does *not* remain from one interview m to another interview n:

$$\begin{aligned} \sum_{m \in \mathcal{P}_i \cap \mathcal{P}_j} x_{mt} + \sum_{n \in \mathcal{P}_i \cap \mathcal{P}_j} x_{n,t+1} - \sum_{o \in \mathcal{P}_i \cap \mathcal{P}_j} \sum_{\substack{\tau = \max(1, t - d_o + 2): \\ d_o > 1}}^{t} y_{o\tau} \geq 2 w_{ijt} \quad \forall (i,j) \in \mathcal{P}, \\ t \in T \backslash \mathcal{F} : |\mathcal{P}_i \cap \mathcal{P}_j| > 1 \wedge i \notin \mathcal{S} \wedge j \notin \mathcal{S}. \end{aligned} \tag{1.11}$$

Similar to the consecutive interviews constraint (1.9), the preserving pairs constraint applies only to non-secretariat interviewers.

1.4.5 Other Interviewer Constraints

We must define that every interviewer can be in at most one interview at a time. However, some interviews will be undertaken partially by the secretariat, which consists of personnel that can undertake interviews when needed at the venue the secretariat is located in. The secretariat consists of homogeneous interviewers who can attend any meeting if needed, and each representative within this secretariat may be attending an interview at time t. Thus, a constraint is needed to specify the maximum number of simultaneous interviews that can be attended. The parameter s_i is equal to one for ordinary interviewers and may be greater than one for the secretariat at the venue. Then:

$$\sum_{m \in \mathcal{P}_i} x_{mt} \leq s_i \quad \forall i \in \mathcal{I} \cup \mathcal{S}, t \in T \tag{1.12}$$

Additionally, the project organizations specified that seniority within groups should be respected. Specifically, for every time slot, the more senior group member must be assigned to undertake an interview before the less senior group member when there are multiple interviewers within a group:

$$\begin{gathered}\sum_{m \in \mathcal{P}_i \cap \mathcal{L}_l} x_{mt} + \sum_{\ell \in L : \ell \neq l} z_{i\ell d} \geq \sum_{n \in \mathcal{P}_j \cap \mathcal{L}_l} x_{nt} \quad \forall t \in T, g \in G \backslash \mathcal{S}, p \in 1, 2, .., (|\mathcal{G}_g| - 1), l \in L, \\ i = member(\mathcal{G}_g, p), j = member(\mathcal{G}_g, p+1), d = member(\mathcal{D}_t, 1) : |\mathcal{G}_g| > 1 \wedge u_{it} = 0\end{gathered} \tag{1.13}$$

The second term on the left-hand side of (1.13) will be equal to one if the more senior interviewer i conducts interviews at a *different* venue than venue l, which will leave the first term on the left-hand side equal to zero. By including both terms on the left-hand side, the seniority concern is respected when the more senior interviewer is present at the venue l and disregarded otherwise. Additionally, the seniority concern is disregarded if the more senior interviewer is unavailable at time t as constraints are defined only if that interviewer is available.

1.4.6 Variable Constraints

$$x_{mt} \in \{0, 1\} \quad \forall m \in M, t \in T \tag{1.14}$$

$$y_{mt} \in \{0, 1\} \quad \forall m \in M, t \in T \tag{1.15}$$

$$z_{ild} \in \{0,1\} \quad \forall i \in \mathcal{I}, l \in L, d \in D \tag{1.16}$$

$$v_{it} \in \{0,1\} \quad \forall i \in \mathcal{I}, t \in T \tag{1.17}$$

$$w_{ijt} \in \{0,1\} \quad \forall (i,j) \in \mathcal{P}, t \in T \tag{1.18}$$

1.4.7 Objective Function

The next step is to define the objective function. With $\omega^A + \omega^C + \omega^P = 1$ representing how much to focus on assigning interviews, achieving consecutive interviews, and achieving preserving interview pairs, respectively, and all at favorable times, the three objectives requested by the student association can be respected at a preferred weighting. While the project organizations primarily favored the two latter objectives, the assignment objective is included in the objective function to ensure that interviews can still be allocated to a favorable time slot, although they cannot necessarily be scheduled consecutively. Then the objective function is defined as:

$$\max \quad \sum_{m \in M} \sum_{t \in \mathcal{B}_m} \omega^A p_t y_{mt} + \sum_{\substack{i \in \mathcal{I} \setminus \mathcal{S}: \\ |\mathcal{P}_i| > 1}} \sum_{t \in T \setminus \mathcal{F}} \omega^C p_t v_{it} + \sum_{\substack{(i,j) \in \mathcal{P}: \\ |\mathcal{P}_i \cap \mathcal{P}_j| > 1 \\ \wedge i \notin \mathcal{S} \wedge j \notin \mathcal{S}}} \sum_{t \in T \setminus \mathcal{F}} \omega^P p_t w_{ijt} \tag{1.19}$$

1.5 Computational Study

Computations were done on a Linux Ubuntu 18.04 computer with an 18-core, hyper-threaded 2.6 GHz Intel Xeon Platinum 8272CL processor and 144 GB RAM. Problems ran with the mathematical programming language AMPL and the commercial solver CPLEX version 12.10.0. CPLEX ran each instance for one hour unless optimality was found earlier.

1.5.1 Selecting the Priority Weights

The model was applied to five test instances from equally many hiring rounds within the project organizations. These instances involve scheduling 17–1149 meetings, and the overview of these instances can be seen in Table 1.4. While the smallest test instance could quickly be scheduled manually, this test instance was nevertheless scheduled by the model to understand how well the model performs on small instances. To select a general set of priority weights $(\omega^A, \omega^C, \omega^P)$, every test instance was run 60 times, differing only by different priority weights $(\omega^A, \omega^C, \omega^P) \in \big\{(i, j, k) : i, j, k \in \{0, 0.1, 0.2, .., 1\} \wedge i + j + k = 1\big\}$. These runs allow us to study how sensitive each objective is to a different tuple of priority weights and in which range we can achieve consistently good results across test instances.

Through experimentation, gradual improvements, and discussions with the project organizations about their preferences, the priority weights $(\omega^A, \omega^C, \omega^P) = (0.3,\ 0.5,\ 0.2)$ were selected. While all three objectives attempt to schedule meetings at favorable times, the assignment objective also assigns meetings at favorable times when consecutive interviews are not achieved. Since this situation frequently occurs for the test instances presented in this article, assigning a strictly positive weight to the assignment objective is valuable. Additionally, in most interview scheduling instances of the size encountered in the project organizations, interviewers usually interview with a varying set of interviewers. This observation makes it challenging to preserve the interviewer pair across consecutive interviews and hence calls for a need to assign the consecutive-interviews objective a large weight. Hence, $(\omega^A, \omega^C, \omega^P) = (0.3, 0.5, 0.2)$ provides consistently good results for all three objectives across all test instances and will be used when studying the test instances. These weights also aligned well with the stated ambitions of the project organizations.

This finding is illustrated with test instance 4 in Figure 1.1. There are three triangles—one for each assigned, consecutive, and preserving objective, respectively. Each triangle shows how well the objective is respected for different weights $(\omega^A, \omega^C, \omega^P)$. Each triangle is scaled relative to the highest attained value for the objective in question—of which a black color in the triangle represents this highest value—to make the triangles more interpretable. The lightest areas in the triangle represent the lowest attained value for the objective in question, except for outlier values that have been modified to improve the interpretability of each figure.

The weights $(\omega^A, \omega^C, \omega^P)$ follow a linear transformation from $\mathbb{R}^3$ to $\mathbb{R}^2$. For example, each corner of the triangle represents a 100% weighting on the labeled objective in the respective corner, such as $(\omega^A, \omega^C, \omega^P) = (1, 0, 0)$ for the lower-left corner. The middle points represents an even weighting of each objective—$(\omega^A, \omega^C, \omega^P) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Furthermore, the sides of the triangles represent a non-zero weight of two objectives and a zero weighting of the third objective. Hence, one can compare how well each objective performs a multitude of $(\omega^A, \omega^C, \omega^P)$ values by locking in the weights $(\omega^A, \omega^C, \omega^P)$—exemplified by the white circle at $(\omega^A, \omega^C, \omega^P) = (0.3, 0.5, 0.2)$ in each triangle—and study the attained relative values of each triangle for those weights.

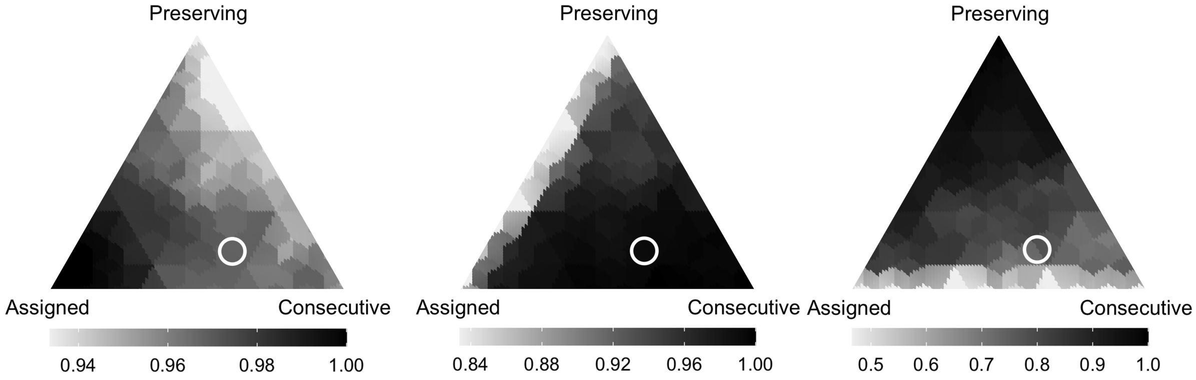

Figure 1.1. Test instance 4's scaled objective values for the assignment, consecutive, and preserving pairs objectives, respectively.

1.5.2 Comparing Time Preference Allocations

Since the time preference parameter is included in all three terms in the objective function, allocating interviews at favourable time slots is naturally a key concern. Hence, it is valuable to construct a single comparable metric that explores how well time preferences are met across test instances. By employing an area under the curve (AUC) score, we can describe how favourable the assignments are.

In a standard AUC curve, a value of one is achieved if all interviews are allocated to the most favoured time slots because the area under this cumulative ratio curve equals one. However, assigning all interviews to the most favoured time slots may not be feasible because there may not be sufficient time slots or rooms available, or if there are too few interviews to assign an interviewer to fill up all time slots. Hence, we must adjust the definition of what would give a full AUC score. Specifically, we must identify the highest number of cumulative assignments that are *practically* possible to allocate and adjust the AUC score accordingly. Ultimately, the modified AUC score is calculated as the area below the cumulative, practically maximum number of assignments, and it is this score that is referred to when mentioning the AUC in the remainder of this section.

1.5.3 Test Instances

Table 1.4. Results for all five test instances employing the priority weights $(\omega^A, \omega^C, \omega^P) = (0.3, 0.5, 0.2)$.

Instance	1	2	3	4	5
Interviewers	56	7	26	21	50
Applicants	812 (807)	17	113	98	1149
Time slots	240	96	96	72	180
Days	5	4	4	3	5
Variables	786 765 (786 765)	6752	60 784	33 978	652 575
Constraints	358 908 (359 001)	5446	28 282	32 357	419 484
Running time (seconds)	Infeasible (3600)	1.00	3600	3600	3600
Relative MIP gap (percent)	Infeasible (37.8)	0	5.38	1.91	53.71
AUC (percent)	Infeasible (85.9)	100.0	97.2	97.7	81.3
Consecutive allocations (percent)	Infeasible (90.6)	94.1	98.1	97.7	84.2
Preserving pairs (percent)	Infeasible (16.6)	8.3	17.5	53.4	9.3

Each test instance varies in size, performance, or feasibility. Specifically, test instance one attempts to schedule 812 interviews, but due to excessive unavailability among the interviewers—with some being unavailable the first four days and their interviewing partner being unavailable the fifth and final days—the model was rendered infeasible. After modifying the model to handle infeasibilities, as described in Section 1.4.1, 807 of these 812 interviews were successfully scheduled. Due to the size of the problem—given the number of interviewers, applicants, and time slots—the relative MIP gap was as high as 37.8% when the stopping time of 3600 seconds was reached. Nevertheless, 90.6% of the meetings were scheduled consecutively, and an AUC score of 85.9% was achieved.

The results are similar to test instance five, in which 1149 interviews must be scheduled. Despite a high relative MIP gap of 53.71%, the highest among all test instances, 84.2% of interviews are scheduled consecutively, and an AUC score of 81.3% is achieved. Despite few interviewer pairs being preserved in consecutive meetings, the quality of this schedule is a substantial improvement over comparable schedules the project organizations produced manually in the past.

The results improve substantially compared to test instances one and five for smaller test instances. Test instance two schedules only 17 interviews and achieves optimality in one second. Test instance three and four, in comparison, does not achieve optimality within the 1-hour running time but achieve low relative MIP gaps of 5.38% and 1.91%, respectively. While both the AUC scores and the number of consecutive allocations improved markedly compared to instances one and five, the most noteworthy improvement is the increase in the percentage of interviewer pairs being preserved across consecutive interviews.

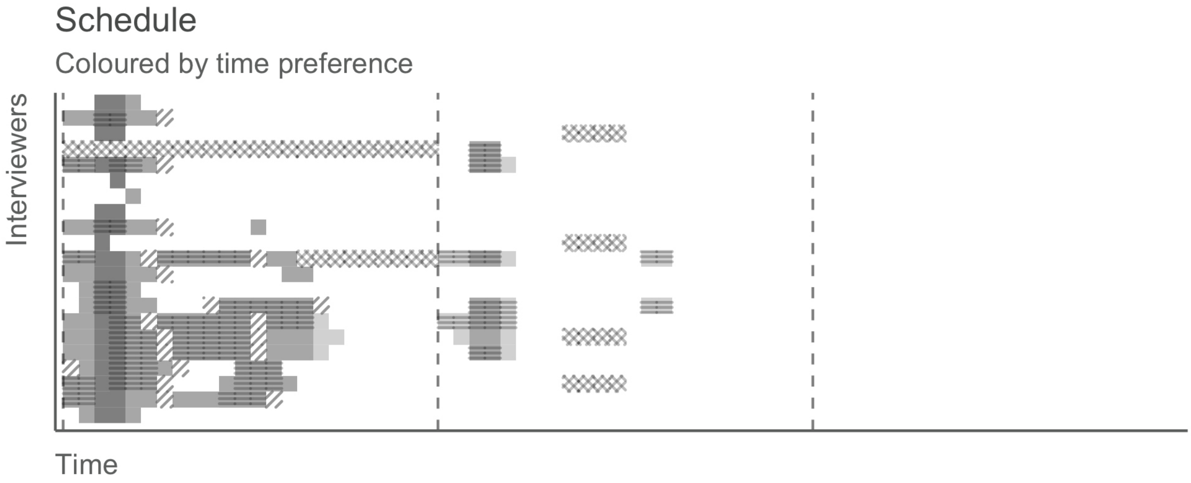

Figure 1.2. Interview schedule for test instance 4.

Figure 1.2 shows the schedule in test instance 4 when the solver was stopped after 60 minutes. The horizontal axis represents the time slots—with the dashed vertical lines being the day separators—and the vertical axis represents the interviewers. The solid blocks represent the assigned interviews, and the shades represent the time preference—with a darker shade symbolizing a stronger time preference. Moreover, the crosshatches represent the time slots in which the interviewers are unavailable, which means that meetings involving these interviewers should not be scheduled for these times. Additionally, angled lines between blocks of consecutive interviews represent the break we enforce when consecutive interviews would otherwise be too long for the interviewers. By contrast, the horizontal lines overlapping the solid blocks represent consecutive interviews in which the interviewer pair was preserved.

Additionally, Figure 1.2 shows that most interviews are allocated at favorable times. However, while all interviews could theoretically have been assigned for the first two time-preference levels, only 87.8% were. The primary reasons for this lower score are the unavailability of two interviewers in particular—which is illustrated with crosshatches in Figure 1.2—and that the model also attempts both to schedule interviews consecutively and preserve pair allocations.

For test instance 4, the scheduling quality is satisfying with an AUC of 97.7%, consecutive allocations for 97.7% of the meetings, and preserving pair allocations for 53.4%. These metrics are substantially better than when similar schedules were produced manually, which is testified by the person responsible for the schedule:

> If one compares the plan's quality with how it would have turned out if we had made it manually, we see that it is much better. We had none of the typical "human" errors we experienced when doing it manually in earlier interviewing rounds, such as double bookings. Instead, the interviewers experienced an excellent flow, with many consecutive interviews early in the week and frequent breaks. That way, we could do interviews for any potential leftover positions at the end of the week. We are incredibly pleased with the produced plan [and the corresponding workflow], as it was efficient, simple to work with, and of good quality.

Moreover, one project organizer noted that the recruiting part of student association work is a key reason students emphasize when explaining why they do not want to engage and volunteer in the student association. These considerations, coupled with the fact that the students are allocated to responsibilities they prefer and the leaders of these responsibilities must conduct these interviews, cause substantial overhead efforts that must be made. Consequently, the student association reports that they have experienced positive effects on their recruiting process due to three main factors.

First, the project organizations have saved substantial time in creating the interview schedule, with the project organizer for test instance one reporting that they reduced their scheduling time from approximately 100 hours to 1 hour. However, while the time savings are large on an individual basis, only a few individuals are involved in creating the schedule. Hence, this factor's impact on the student association is only modest.

In comparison, the second factor does not save any meeting time on an individual basis because the number and duration of interviews are pre-defined, but utilizes the interviewers' time at large more efficiently. Specifically, by scheduling numerous consecutive interviews, the interviewers have more time to study without frequent interruptions to conduct interviews. Both since frequent interruptions were common in the manually produced schedules and since many interviewers are affected by a better schedule, the schedule's impact on the student association has been reported by the project organizations as substantial.

And third, while the project organizations must hire approximately 3000 volunteer students over a two-year duration, many of these students have previously participated in the project organizations but usually in different positions. Therefore, many students have already experienced first-hand how relatively frictionless interviewing rounds were when schedules were produced with the model presented in this article. Consequently, the project organizations have reported that it is generally easier to recruit students because the time spent on recruiting is not as big a concern as it originally was when the schedules were produced manually. Each schedule produced by the model in this article hence creates positive spillover effects, which hopefully will contribute to sustaining a high volunteering volume in these project organizations.

Additionally, the second aim of this article is to increase awareness of and enthusiasm about operations research in general among the student mass. Across these 5 test instances, 160 interviewers and approximately 30 secretariat members have experienced the direct benefits of automated scheduling and improved meeting schedules compared to former, manually created schedules. Throughout multiple interviews with interviewers and the secretariat, it is clear that the second aim has been achieved.

Ultimately, one main aim of these project organizations is to provide future leaders with sought-after work experience. As many of the students participating in the project organizations will take on future leading positions in the business landscape and executives believe leadership skills improve with volunteering (Deloitte, 2016), it is the hope that the impact of this decision support tool will extend well beyond its direct effects for the students at NHH and that the increased awareness of OR tools will increase its spread and usage.

1.6 Concluding Remarks

This article provides an applied model to schedule en-masse hiring rounds within volunteer organizations. An integer linear programming model with three weighted objectives was applied to this problem: to assign interviews at favorable times, assign interviewers with consecutive interviews at favorable times, and assign consecutive interviews at favorable times in which the interviewer pairs are preserved.

The weighting of each objective was first assigned with various values, and advantageous weights of each objective were identified. Due to the extensively intersecting properties of the three objectives, the generally most advantageous objective function value was achieved with a high weighting of the preserving pairs objective, a small but strictly positive weighting of the consecutive pairs objective, and an even smaller yet strictly positive weighting of the simple assignment objective.

The model has been implemented in the hiring rounds of the student association at NHH, studying test instances ranging from 17 to 1149 meetings to be scheduled. Compared to former, manually created schedules, both the volunteer organization's and the interviewers' perceived experience is markedly improved when a non-trivial number of meetings are scheduled, with considerable improvements in all three model objectives. Due to the high degree of satisfaction after implementing the model, the student association intends to continue using the model for each future hiring round.

The model presented in this article can be expanded to further contribute to the literature in two primary directions. First, future work should incorporate a fair allocation of meetings to make the individuals' schedules more evenly good. This feature is missing from the literature altogether today. Second, during COVID-19, whole organizations moved to online communication platforms and digitized their work calendars across the organization. Since meetings create attention residue for many office workers, making it particularly difficult to focus on cognitively demanding tasks shortly after such events (Leroy, 2009), it is valuable to schedule these meetings more efficiently—such as more consecutively. Therefore, future work also involves modifying the meeting scheduling problem to produce updated schedules frequently to let workers utilize their working days better.

References

Aizam, N. A. H., & Sim, S. L. P. (2016). Integrating mathematical model and scheduling problem. *AIP Conference Proceedings*, *1775*, 030065.

Bartholdi III, J. J., & McCroan, K. L. (1990). Scheduling interviews for a job fair. *Operations Research*, *38*(6), 951–960.

Corporation for National and Community Service. (2013, June). Volunteering as a pathway to employment: Does volunteering increase odds of finding a job for the out of work? https://americorps.gov/sites/default/files/%5C%5Cevidenceexchange/FR_2013_VolunteeringasaPathwaytoEmployment_1.pdf

Deloitte. (2016, June). Building leadership skills through volunteerism.

Ernst, A. T., Mills, R., & Welgama, P. (2003). Scheduling appointments at trade events for the Australian Tourist Commission. *Interfaces*, *33*(3), 12–23.

Gebser, M., Glase, T., Sabuncu, O., & Schaub, T. (2013). Matchmaking with answer set programming. *International Conference on Logic Programming and Nonmonotonic Reasoning*, 342–347.

Gordon, L., & Erkut, E. (2004). Improving volunteer scheduling for the Edmonton Folk Festival. *Interfaces*, *34*(5), 367–376.

Gueret, C., Bellenguez-Morineau, O., Pavageau, C., Péton, O., & Poncelet, D. (2009). Optimisation de la planification de bourses d'échanges de technologies. *ROADEF*, *9*, 177–178.

Huang, Y., Zhang, X., & Alexander, P. S. (2012). A heuristic algorithm for optimizing business matchmaking scheduling. *International Journal of Operations Research and Information Systems (IJORIS)*, *3*(4), 59–73.

Kiyonari, Y., Miyano, E., & Miyazaki, S. (2006). Computational complexity issues in university interview timetabling. *Practice and Theory of Automated Timetabling*, *6*, 448.

Le Roux, A., Bellenguez-Morineau, O., Guéret, C., & Prot, D. (2015). Complexity of a one-to-one meeting scheduling problem with two types of actors. *Journal of the Operational Research Society*, *66*(10), 1730–1741.

Leroy, S. (2009). Why is it so hard to do my work? The challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes*, *109*(2), 168–181.

Mausser, H. E., Magazine, M. J., & Moore, J. B. (1996). Application of an annealed neural network to a timetabling problem. *INFORMS Journal on Computing*, *8*(2), 103–117.

Pesant, G., Rix, G., & Rousseau, L.-M. (2015). A comparative study of MIP and CP formulations for the B2B scheduling optimization problem. *International Conference on AI and OR Techniques in Constriant Programming for Combinatorial Optimization Problems*, 306–321.

Rinaldi, F., & Serafini, P. (2006). Scheduling school meetings. *International Conference on the Practice and Theory of Automated Timetabling*, 280–293.

Sampson, S. E. (2006). Optimization of volunteer labor assignments. *Journal of Operations Management*, *24*(4), 363–377.

Schrage, L. Scheduling job placement interviews at a university [PATAT 2004 Conference]. In: PATAT 2004 Conference. PATAT. 2004.

Shin, S., & Kleiner, B. H. (2003). How to manage unpaid volunteers in organisations. *Management Research News.*

United Nations Volunteers programme. (2018). The thread that binds–volunteerism and community resilience.

Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., & De Boeck, L. (2013). Personnel scheduling: A literature review. *European journal of operational research*, *226*(3), 367–385.

Assessing the Impact of Driver Overtime in the Distribution Network of a Flower Retail Chain

Christian Braathen and Mario Guajardo

Abstract

This article studies the impact of social constraints on the vehicle routing problem, with a particular focus on allowing overtimes for the drivers. Working overtime is common in practice, as it may improve driver utilization, but it also requires a more detailed cost structure in the routing problem. We formulate a mixed integer linear model for this problem, motivated by an application at a florist company performing daily routes in a network of stores in Norway. The model considers a single-depot, multi-trip situation consisting of a heterogeneous fleet of capacitated trucks with time windows, deliveries and split pickups, asymmetric distances, and real-life speed limits. By emphasizing and incorporating social constraints both in the model and in a route-generating process, the results outperform manually produced schedules and a commercial software, with cost reductions totaling 17.4%–36.4% and 9.7%–25.5%, respectively, while achieving lower solving times. We also run the model with a traditional distance minimization perspective, and show how the routes and cost obtained differ from when the overtimes are explicitly included in the problem. Notably, the results illustrate that overtimes are beneficial for cost savings and they are most valuable to serve locations far away from the headquarters.

Keywords: vehicle routing problem (VRP); assignment problem; social constraints; mixed integer linear programming

JEL Classification: **J22** Time Allocation and Labor Supply; **R41** Transportation Demand; **L81** Retail and Wholesale Trade; **C61** Optimization Techniques; **C44** Operations Research

2.1 Introduction

While the vehicle routing problem (VRP) has been thoroughly studied, a large share of studies ignore social constraints (Lahyani et al., 2015). These constraints, which relate to the drivers, are important for several reasons. First, driver safety is a real-life concern and should be considered for practical considerations in the VRP. Second, implementing social constraints impacts total transportation costs, total driving distance—and other commonly used objective function values—due to their constraining nature. Third, social considerations may be substantially impacted by each type of objective function unless constrained sufficiently. Hence, we need to understand to which extent these considerations are impacted.

Motivated by an application at the largest wholly-owner florist chain in Norway, this article addresses the question of how social constraints affect the routes and the quality of the solution. We are particularly interested in studying the effect of allowing drivers to work overtime, as this is a usual practice but commonly not included in the formulation of the routing problem. For this purpose, we formulate a mixed integer linear programming model, characterized by a heterogeneous fleet of capacitated trucks, multi-trips, time windows, deliveries and split pickups, asymmetric distances, real-life speed limits, and driver availability. To explore the effect of social constraints on different metrics, we implement different variants of the model, including objective functions with and without the consideration of overtimes.

Although the concepts of social constraints are acknowledged by the practitioners in charge of routing, it can be challenging to grasp all the details of the routes, the trade-offs between them, and the implications of social constraints for the total costs. In practice, planners tend to manually tweak the software solutions to cope with these social constraints or manually do the whole routes from scratch. With access to real data instances and both the practitioners' and commercial software's solutions, we have studied in detail the impact of social constraints in this VRP.

The remainder of this article is organized as follows. Section 2.2 explores the VRP literature in the context of social constraints. Section 2.3 describes the mathematical formulation of the model, while Section 2.4 performs a computational study and discusses key insights obtained from the results. The article concludes in Section 2.5.

2.2 Literature review

The Vehicle Routing Problem (VRP) represents an essential challenge in transporting goods and services, with its fundamental objective being to identify an optimal set of routes that service a set of customer demands (Adewumi & Adeleke, 2018). Although exact methods provide optimal solutions, they are best suited for smaller-scale problems due to the trade-off between solving time and solution quality. By comparison, heuristic and meta-heuristic methods can handle more extensive problems and a variety of constraints, delivering near-optimal solutions. With an expanding array of problem variants and potential solutions to complex real-world attributes (Eksioglu et al., 2009), in addition to stronger computational power, the study of the VRP has witnessed a remarkable evolution at a 6% annual growth rate (Tan & Yeh, 2021).

The most common objective of real-world VRP variants is to minimize total distribution costs while upholding high-quality distribution services (Konstantakopoulos et al., 2020). The total distribution cost constitutes fixed and variable costs and a hybrid between the two. Fixed costs mainly stem from fixed driver wages and fixed vehicle usage costs. In contrast, variable costs are primarily driven by fuel expenses or travel times—including variable driver wages—for each route (Konstantakopoulos et al., 2020). The hybrid costs, by comparison, include toll road costs that the vehicle pays for once but has an unlimited number of passes within a time window—such as one hour—but must pay anew once the time window expires. Factors such as hybrid costs extend cost considerations beyond the mere calculation of the travel expenses and increase the complexity of the problem (Reinhardt et al., 2016). The route's length and duration, influenced by constraints and parameters of real-world routing problems, are essential when seeking the most cost-effective optimization method to solve a particular VRP variant (Konstantakopoulos et al., 2020).

2.2.1 Driver considerations

While driver wages are a substantial component of total distribution costs—the most common objective in VRP—drivers receive little focus in VRP studies. Specifically, in the taxonomy of Tan and Yeh (2021), only two components explicitly mention drivers and only a minority of VRP articles include working hour considerations and legislation protecting the drivers—such as required breaks when driving throughout the day (Lahyani et al., 2015).

Furthermore, the concept of working duration might be ambiguous, and subject to specific stipulations from local legislations. Hence, Rincon-Garcia et al. (2020) proposed the following four general definitions not involving regional legislation. First, “route duration” refers to the period a driver and its vehicle spends after departing the depot until returning to the depot after the last service request. Second, “accumulated working time” is the accumulated time during which the driver performs tasks, such as driving, loading, and unloading, and predicted or unpredicted waiting times—for example, due to delays or arrival prior to time windows. Third, “accumulated driving time” describes the accumulated time driven by the vehicle prior to having a required break from driving. Accumulated driving time is reset after each such break. Fourth, “total accumulated driving time” is the driver’s aggregated driving time throughout the working day. Each of these terms may have its own set of constraints in VRP.

2.2.2 Overtime considerations

A common approach for urban problems is to assume no overtime is needed to serve the customer furthest away from the depot during an ordinary shift, which is the approach of Jula et al. (2005). If this assumption is coupled with an assumption of a fixed salary for permanently hired drivers, then the entire driver wage costs are irrelevant for the cost minimization problem (Horng & Yenradee, 2020). However, the no-overtime assumption is unfavorable because a trade-off between increased overtime costs—which typically comes at a higher salary rate—and reduced other operational costs may be beneficial and lead to a more efficient solution (Ren et al., 2010). This finding is demonstrated by Simeonova et al. (2018), who identified that allowing overtime yielded a 12% improved driver utilization in terms of working hours, in addition to a 12.5% improved vehicle utilization without a substantial increase in costs.

While overtime is straightforward to calculate for single-trip models—which most VRP studies assume (Karoonsoontawong et al., 2020)—with fixed delivery and pickup amounts, calculating overtime is more complicated for multi-trip models. Specifically, the former models can incorporate the overtime cost directly into the cost parameter. In contrast, multi-trip models and models with split deliveries or pickups must incorporate overtime as a decision variable. The latter is the approach of Ren et al. (2010). Moreover, the model must register overtime work only after the regular working hours have been scheduled (Simeonova et al., 2018), which complicates the solving process further. For a comprehensive overview of multi-trip models, see Cattaruzza et al. (2016).

When allowing overtime, there are commonly three maximum limits of overtime being used. Two approaches are limiting overtime usage per driver by the company’s policy or by legislation

(Ren et al., 2010). However, a third approach is to minimize the amount of overtime used via the objective function. This can be done using two different procedures, as developed in Ouaddi et al. (2018). A first one consists of directly minimizing the maximum total overtime used by the vehicle fleet. The alternative procedure consists of first identifying the lowest level of maximum overtime needed for any individual driver, to create a feasible solution and place this limit as a constraint in a second problem. While the first procedure creates more cost-efficient solutions, the latter respects driver wellness more by creating a more equitable and sustainable solution.

Analyzing overtime can also be used to consider tactical considerations. One natural approach is to analyze whether the number of drivers can be reduced, which is part of the analysis for this article. A second approach is changing a distribution company's level of outsourcing when overtime is allowed, which Moon et al. (2012) explored. A third approach is evaluating overtime usage with other decisions, such as changing the vehicle fleet size. Simeonova et al. (2018) explored this third approach and identified potential cost savings of 8% with this trade-off when allowing a maximum overtime usage of 4.5 hours per driver. A fourth approach is to identify how much overtime to allow to preserve driver relationships with the distribution company's customers, which Horng and Yenradee (2020) explored.

Generally, a lack of focus on social constraints is unfavorable for three primary reasons. First, the distribution company's costs may not reach an optimal level when ignoring a key cost component. Second, there is a societal downside as the drivers' fatigue is a significant factor for 15–20% of collisions involving commercial vehicles (Goel & Vidal, 2014). Third, with the number of problem variants in the VRP being large and still growing, inattention to real-world concerns about drivers limits our understanding of new problem variants and the impact that social constraints may have. Moreover, it would be practical to evaluate the impact of social constraints across different VRP problem variants tested on the same test instances to broaden the understanding precisely. This article aims to study such different driver considerations and evaluate their impact across different common objective functions in a real-life context.

2.3 Mathematical Model

The mixed integer linear programming model presented in this section aims to study social constraints to broaden the understanding of how they impact routing when different constraints and objective functions are applied. In order to have a realistic assessment of social constraints, the model considers real-life properties such as capacity constraints, a single depot, time windows, deliveries, split pickups, asymmetric distances, and speed limits. All of these properties, except for pickups and multi-trips, are incorporated into a route-generating process, which is described in detail in Appendix 2.A. Then, the candidate routes are fed into a route-based model formulation, which also incorporates pickups and multi-trips. The model outcome allocates drivers and vehicles to routes, and the pickup amounts on each allocated route, while taking into account the various social constraints.

2.3.1 Notation

Table 2.1. Model components.

Sets	
D	Set of drivers.
V	Set of vehicles.
R	Set of routes.
S	Set of stores.
P	Set of pickup locations.
Subsets	
$\mathcal{O} \subset (D, V, R)$	Set of assignable (driver, vehicle, routes) alternatives.
$\mathcal{P} \subset (D, V, R, P)$	Set of assignable (driver, vehicle, routes, pickup locations) alternatives.
$\mathcal{O}^S_s \subset \mathcal{O}$	Set of assignable (driver, vehicle, routes) alternatives in which store s is in the route.
$\mathcal{O}^D_d \subset \mathcal{O}$	Set of (driver, vehicle, routes) alternatives in which the driver is driver d.
$\mathcal{P}^R_{dvr} \subset P$	Ordered set of pickup locations on the route.
Parameters	
$\alpha^r_{dvr} \geq 0$	Route costs excluding overtime cost and pickup costs.
$\alpha^o_d \geq 0$	Overtime cost per minute.
$\rho^p_{dvrp} \geq 0$	Pickup costs per container.
$\rho^t_p \geq 0$	Pickup time per container.
$\rho^d_p \geq 0$	Pickup demand.
$\rho^c_{dvrp} \geq 0$	Cumulative pickup capacity.
$\omega \geq 0$	The max allowable hours in a working day, including overtime.
$\omega^n \geq 0$	The duration of a normal working day, excluding overtime.
$\gamma^a \geq 0$	A problem variant-dependent limit on longest working day, including overtime.
$\gamma^r \geq 0$	The highest allowable number of driving minutes before needing a break (4.5 hours).
$\gamma^b \geq 0$	The minimum break duration in minutes (45 minutes).
$\delta^r_{dvr} \geq 0$	Route duration.
$\delta^d_r \geq 0$	Driving time.
$\beta_r \geq 0$	Route r's distance.
$\epsilon \geq 0$	A minuscule small amount.
Variables	
$x_{dvr} \in \{0, 1\}$	Driver d uses vehicle v to drive route r.
$y_{dvrp} \in \mathbb{N}_0$	The amount of pickup containers picked up from location $p \in P$ on route r with (driver,vehicle)= (d, v).
$z^o_d \geq 0$	The amount of overtime the driver d has during the workday.
$z^w_d \geq 0$	The amount of time the driver d works during the workday, driving between locations, delivering containers, and picking up containers.
$z^d_d \geq 0$	The total amount of driving that driver d is driving during the workday.
$z^r_d \geq 0$	The number of minutes the driver d must rest given the workday's driving amount.
$z^b_d \in \mathbb{N}_0$	The number of 45-minute blocks driver d must rest.
$v \geq 0$	The minimax working day—a variable for problem variant A and a parameter otherwise.
$w_d \in \{0, 1\}$	Driver d is assigned routes.

2.3.2 Constraints

Pickup and Delivery Constraints

$$\sum_{(d,v,r)\in\mathcal{O}^S_s} x_{dvr} = 1 \quad \forall s \in S \tag{2.1}$$

$$\sum_{(d,v,r,p)\in\mathcal{P}} y_{dvrp} = \rho^d_p \quad \forall p \in P \tag{2.2}$$

$$\sum_{\substack{p'\in\mathcal{P}^R:\\ ord(p')\leq ord(p)}} y_{dvrp'} \leq \rho^c_{dvrp} \quad \forall (d,v,r) \in \mathcal{O}, p \in \mathcal{P}^R_{dvrp} \tag{2.3}$$

$$x_{dvr} \geq \frac{1}{\rho^d_p} y_{dvrp} \quad \forall (d,v,r,p) \in \mathcal{P} \tag{2.4}$$

Constraints (2.1)–(2.4) specify the delivery and pickup constraints. First is (2.1), which states that every store with a delivery order must be visited exactly once. While this constraint could potentially have caused infeasibility issues if a store places a large order, it has no practical infeasibility problems for the test instances in this article. Additionally, while this constraint might also cause an inefficient utilization of the vehicles compared to a multiple-delivery approach, the added value from multiple deliveries are usually too low to justify it due to the proximity of most locations. Hence, this trade-off favors only one delivery per location.

Moreover, there are three pickup constraints (2.2)–(2.4). Constraint (2.2) states that all pickup points' demand must be covered precisely regarding the number of carts that need to be collected. While this constraint may be relaxed in specific applications, it is essential to collect all carts intraday because the goods in this application are fresh.

When collecting carts from pickup locations, the vehicle's practical capacity ρ^c_{dvrp} must always be respected, which (2.3) ensures. The practical capacity is defined as the vehicle's actual capacity minus the number of carts yet to be delivered on route r *and* the additional empty space required. Additionally, (2.4) states that if the (driver, vehicle, route) permutation (d,v,r) is allocated—i.e., x_{dvr} equals 1—then pickups can occur for this permutation and is disallowed otherwise. Finally, this constraint states that if pickups are enabled, the pickup amount will maximally be equal to the demanded amount of carts to be picked up.

Resting Requirement Constraints

$$\sum_{(d,v,r)\in\mathcal{O}_d^D} \delta_r^d x_{dvr} = z_d^d \quad \forall d \in D \tag{2.5}$$

$$z_d^d \geq \gamma^r \cdot z_d^b - \epsilon \quad \forall d \in D \tag{2.6}$$

$$z_d^d \leq \gamma^r (z_d^b + 1) - \epsilon \quad \forall d \in D \tag{2.7}$$

$$z_d^r = \gamma^b z_d^b \quad \forall d \in D \tag{2.8}$$

Constraints (2.5)–(2.8) specifies the resting requirements of each driver d. First, constraint (2.5) calculates the total time driver d is driving a vehicle during the day. Using this variable, constraints (2.6) and (2.7) defines the floor and ceiling of the number of 45-minute resting breaks z_d^b that driver d should have, respectively. With the decision variable's integer constraint (2.20) on page 36, the number of rest breaks will be the integer value that falls between the floor and the ceiling values. Each of these breaks must be 45 minutes long, and (2.8) calculates the total required resting time driver d faces during the day.

Working Time Constraints

$$\sum_{(d',v,r)\in\mathcal{O}_i^D} \delta_{d'vr}^r x_{d'vr} + \sum_{(d',v,r,p)\in\mathcal{P}} \rho_p^t y_{d'vrp} = z_d^w \quad \forall d \in D \tag{2.9}$$

$$z_d^w + z_d^r \leq v \quad \forall d \in D \tag{2.10}$$

$$z_d^w + z_d^r \leq \gamma^a \quad \forall d \in D \tag{2.11}$$

Constraints (2.9)–(2.11) study driver d's total working time. First, (2.9) specifies that the driver's total working time during the day, z_d^w, is calculated as the time working on routes plus additional time picking up carts at pickup locations because that time is not factored into $\delta_{d'vr}^r$. Next, z_d^w is then applied to (2.10), which is used only for problem variant A and defines the longest working day v worked by the longest-working driver. z_d^w is also applied to (2.11), which specifies that no drivers can work longer during the day, including required rests and allowable overtime, than a prespecified limit. For problem variants A, B, and F, this limit is set to 15 hours. For problem variant C, the limit is set to the duration of the longest working day of the manual schedule. And for problem variant D and E, the limit is set to v.

Deciding how much overtime to allow via γ^a may create a substantial impact because it may directly affect delivery decisions for the costliest and longest routes. For example, while the assumption of no needed overtime may hold, there might exist a cluster of customers the furthest away from the depot that, independently, can be visited within an ordinary shift duration without overtime, but visiting many customers in this cluster is infeasible due to the no-overtime constraint. However, if capacity and time window constraints hold, allowing even some degree of overtime might reduce the number of vehicles needed to serve this cluster, ultimately impacting total distribution costs due to the driving distance between the depot and the cluster.

Overtime Constraint

$$z_d^w + z_d^r - \omega^n \leq z_d^o \quad \forall d \in D \tag{2.12}$$

Constraint (2.12) calculates the total amount of overtime driver d works during the day. Together with the decision variable's non-negativity constraint (2.16), z_d^o will either be zero or the difference between total working time z_d^w including required rest z_d^r and the normal working day's duration ω^n—whichever is the highest.

Driver Usage Constraint

$$\sum_{(d,v,r)\in\mathcal{O}_i^D} x_{dvr} \leq |\mathcal{O}_d^D| w_d \quad \forall d \in D \tag{2.13}$$

The objective of problem variant E is to minimize the number of drivers being used. Hence, an explicit definition must be made, so (2.13) defines when a driver is used (conversely, since the objective function of problem variant E is a minimization problem, there is no need to explicitly define when driver d is not used).

Domain of the Variables

We complete the formulation stating the domain of the variables in constraints (2.14)–(2.22) below.

$$x_{dvr} \in \{0, 1, \} \quad \forall d \in D, v \in V, r \in R \tag{2.14}$$

$$y_{dvrp} \in \mathbb{N}_0 \forall \quad d \in D, v \in V, r \in R, p \in P \tag{2.15}$$

$$z_d^o \geq 0 \quad \forall d \in D \tag{2.16}$$

$$z_d^w \geq 0 \quad \forall d \in D \tag{2.17}$$

$$z_d^d \geq 0 \quad \forall d \in D \tag{2.18}$$

$$z_d^r \geq 0 \quad \forall d \in D \tag{2.19}$$

$$z_d^b \in \mathbb{N}_0 \quad \forall d \in D \tag{2.20}$$

$$v \geq 0 \quad \forall d \in D \tag{2.21}$$

$$w_d \in \{0,1\} \quad \forall d \in D \tag{2.22}$$

2.3.3 Objective Function

Problem variant A is a pre-problem that attempts to identify the minimax working day duration—that is, the shortest working day for the driver who works the longest given the constraints of the problem.

$$\min v \tag{2.23}$$

Problem variant B–D attempts to minimize the total cost, which includes route-driving costs for ordinary working hours, pickup costs for the carts that are picked up, and overtime costs.

$$\min \sum_{(d,v,r)\in\mathcal{O}} \alpha_{dvr}^r x_{dvr} + \sum_{(d,v,r,p)\in\mathcal{P}} \rho_{dvrp}^p y_{dvrp} + \sum_{d\in D} \alpha_d^o z_d^o \tag{2.24}$$

Problem variant E attempts to minimize the number of drivers being used.

$$\min \sum_{d \in D} w_d \tag{2.25}$$

Finally, problem variant F attempts to minimize the total distance driven.

$$\min \sum_{(d,v,r) \in \mathcal{O}} \beta_r x_{dvr} \tag{2.26}$$

Table 2.2 illustrates which objective functions and constraints are utilized across the various problem variants. Constraint (2.11) is particularly interesting as it operates with different working day lengths. Problem variants A, B, and F has no (N) practical limit but uses a fallback limit of 15 hours. Problem variant C is limited by a prespecified max limit (M) that equals the duration of the longest working day in the manually produced schedule. And for problem variant D and E, the longest allowable working day is decided by the pre-problem (P) solution of problem variant A.

Table 2.2. Problem variants.

Problem Variant	*A*	*B*	*C*	*D*	*E*	*F*
Objective	Minimize the longest working day	Minimize total cost	Minimize total cost	Minimize total cost	Minimize drivers used	Minimize total distance driven
(2.1)–(2.9)	Yes	Yes	Yes	Yes	Yes	Yes
(2.10)	Yes	No	No	No	No	No
(2.11)	N	N	M	P	P	N
(2.12)	No	Yes	Yes	Yes	Yes	Yes
(2.13)	No	No	No	No	Yes	No

2.3.4 Generating Candidate Routes

In general, VRPs are hard to solve to optimality (Lenstra & Kan, 1981), and heuristic approaches are often adopted to find good solutions within a reasonable time. Our problem is not an exception, and the model formulation above proves hard to solve for realistic instances, as it is shown in Section 2.4. Therefore, we adopt a route generation approach, which is a common approach in the design of heuristics for VRPs (Braekers et al., 2016). In particular, we generate the routes based on the principle that after visiting a particular location, a vehicle can only visit locations that belong to a set of "neighbors". This set only contains locations with intraday deliveries or pickups. Generating this set consists of five primary steps—choosing locations with relative proximity between locations and overlapping time windows, replacing neighbors with parallel pathways to diversify driving directors, adding locations along the trajectory back to the warehouse, and utilizing the planner's expertise by allowing hardcoded locations to be added to the neighbor set.

After defining the neighbor sets, routes are generated, and route details are added. In the route generation, routes of length i are expanded from routes of length $i - 1$ until a stopping rule is met. These rules include time limits, route set cardinality limits, and vehicle capacity constraints. After the route generation is complete, specific route details are produced. These details include driving schedules, cart load changes, driving, loading, and unloading times, and distances. The Google Maps API is utilized to estimate travel times and distances—which is then used to generate driver and vehicle costs—and data on toll stations are used to generate toll station costs. See Section 2.A for more information about the route-generating process.

2.4 Computational Study

The calculations were executed on a virtual machine running Linux Ubuntu 18.04, powered by an Intel Xeon Platinum 8272CL hyper-threaded processor with 18 cores and a clock speed of 2.6 GHz. The machine was equipped with 144 GB of RAM. The mathematical programming environment AMPL and the commercial optimization solver CPLEX, version 12.10.0, were used to run the optimization problems.

This study has seven test instances, and this limited number of test instances is a deliberate choice beyond just the data availability for the manual scheduling alternative and the commercial alternative that the results are compared against per test instance. Specifically, since this article seeks to enhance the understanding of social constraints in real-life contexts, it is more favorable to study a handful of test instances thoroughly than to study a vast set of test instances.

Furthermore, to understand social constraints, the test instances are assessed using multiple objective functions, including minimizing total costs, minimizing the number of drivers used, and minimizing the total distance driven within each delivery day. We organize the presentation of results in different sections—first with a broad overview of results to each test instance in Section 2.4.1, and then with a deeper study on test instance 1 in the subsequent sections.

2.4.1 Overview of Results

We begin the discussion of results by analyzing the cost minimization problem with the longest working hours limit—problem variant B.

Table 2.3. Overview of test instances.

Instance	1	2	3	4	5	6	7
Locations	52	52	52	41	58	79	78
Drivers	13	13	13	14	16	16	16
Vehicles	14	14	14	15	17	17	17
Routes	25 275	32 353	121 749	20 594	178 218	50 047	10 358
Variables	120 541	147 531	510 681	100 790	808 689	244 595	55 817
Constraints	195	195	195	195	234	255	254
Running time (seconds)	462	58	400	352	1500	1228	54
Relative MIP gap (percent)	1.00	1.00	1.00	1.00	3.64	1.00	1.00

Table 2.3 illustrates features about the data and the solutions of the seven test instances presented in this article. In the case of the florist chain, each delivery day involved 41–79 locations, 13–16 drivers, and 14–17 vehicles. There is, however, not a strong positive correlation between the number of routes and variables and the number of locations, drivers, and vehicles. This lack of correlation occurs because the number of carts to be delivered or picked up at each location varies substantially for each delivery day, which subsequently affects the route-generating process and, ultimately, the number of variables for the test instance.

Table 2.4. Test instance costs.

Instance	1	2	3	4	5	6	7
Total day costs (NOK)							
Our approach	41 556	40 077	30 931	31 790	41 953	66 822	80 074
Manual planning	55 234	51 904	47 924	50 019	57 622	80 893	97 379
Commercial solution		44 394	38 176	37 480	56 276	75 091	93 361
Total driving costs (NOK)							
Our approach	12 223	12 656	12 433	14 050	16 188	25 529	30 156
Manual planning	19 741	19 775	19 775	19 100	23 149	29 890	31 413
Commercial solution		17 742	17 742	16 093	23 751	27 680	30 724
Total salary costs (NOK)							
Our approach	27 962	27 421	18 498	16 527	24 000	38 457	46 845
Manual planning	32 334	32 128	28 147	28 547	31 152	46 633	61 157
Commercial solution		26 651	20 433	19 932	28 556	43 836	58 520
Total toll station costs (NOK)							
Our approach	1371	0	0	1212	1765	2836	3073
Manual planning	3159	0	0	2370	3323	4372	4813
Commercial solution		0	0	1457	3970	3572	4118

The results presented in Table 2.4 illustrate the efficiency of the solving process compared to the manual and commercial alternatives. Specifically, our approach consistently outperforms manual planning and the commercial solution regarding total day costs across all test instances. Compared to the manually produced schedule, our solution outperforms it by at least 17.4%–36.4%, with more considerable relative cost savings achieved for the smallest test instances. However, nominal cost savings are the greatest for the largest test instances. Similarly, our solution outperforms the commercial solution across all six test instances we could obtain data for, with total day cost savings being at least 9.7%–25.5%. These cost savings are substantial for total driving and salary costs, indicating that our approach chooses fewer or shorter routes, less expensive vehicles, and less expensive drivers in terms of average hourly wage rate or total hours worked. These details will be explored next.

Table 2.5. Driving details for test instances.

Instance	1	2	3	4	5	6	7
Routes driven							
Our approach	13	11	10	10	13	21	27
Manual schedule	17	17	17	13	23	25	31
Commercial solution		13	13	9	16	21	24
Total km							
Our approach	1656	1641	1617	1851	2136	3131	3674
Manual schedule	2220	2226	2226	2140	2565	3445	3563
Commercial solution		2154	2154	1960	2829	3165	3572
Total km excl. to/from warehouse							
Our approach	497	547	574	471	724	739	523
Manual schedule	568	575	575	652	556	900	833
Commercial solution		801	801	750	1075	1349	1305

Table 2.5 contributes to understanding the cost savings from our solution, with two primary reasons emerging. On the one hand, our solution generally assigns fewer routes to deliver the same amount of carts to each location, which naturally contributes to reducing total distance driven since locations are dispersed, and fewer routes result in fewer trips to and from the warehouse. However, Table 2.5 also indicates that the total driving distance between delivery and pickup locations— excluding driving to and from the warehouse—selected by our approach largely outperforms the manual schedule and the commercial solution.

Table 2.6. Driver details for test instances.

Instance	1	2	3	4	5	6	7
Drivers used							
Our approach	11	10	10	10	13	16	16
Manual schedule	13	13	13	13	15	17	17
Commercial solution		13	13	9	16	17	17
Total work time (hours)							
Our approach	79.6	67.5	65.5	64.6	84.0	128.4	144.1
Manual schedule	84.2	81.7	81.7	76.0	93.3	134.6	161.1
Commercial solution		73.9	73.9	68.7	105.4	136.3	159.5
Total overtime (hours)							
Our approach	6.3	4.4	4.6	0.9	6.8	15.2	27.0
Manual schedule	3.8	3.4	3.4	7.9	3.3	14.2	34.0
Commercial solution		2.7	2.7	4.1	3.4	21.4	45.3
Average hourly wage rate (NOK)							
Our approach	351.4	406.0	282.3	255.8	285.7	299.6	325.0
Manual schedule	384.1	393.4	344.7	371.0	334.0	346.5	379.7
Commercial solution		360.8	276.6	290.3	271.0	321.6	366.9

Furthermore, cost savings were generally substantial in our solution due to reduced salary costs, which Table 2.6 explores in detail. Across all test instances, the total work time was lower in our solution compared to the manually produced schedule and the commercial solution. While the reduced work time occurs partially because fewer drivers were generally utilized in our solution, the reduced number of drivers also contributes to greater overtime use for four test instances. In addition to the reduced total working time, our approach has also selected the most cost-efficient drivers who, including overtime payments, contribute to generally reducing the average hourly wage rate for the utilized drivers.

While the average overtime usage across solving methods is negligibly different, the variance of the drivers' overtime within the test instances is the lowest in our solution. When studying overtime usage, variance among drivers is a valuable metric to evaluate because it signals how well the overtime is distributed among the drivers and can hence be considered an indication of how fair the schedule is from an overtime perspective. Moreover, this metric is particularly useful if test instances employ substantial overtime, and the higher variance in the other solving methods is attributed mainly to these test instances.

Another valuable metric that indicates how taxing a solving method's solution can be, is to study how high the average overtime per driver becomes across test instances. This maximum average overtime per driver per day—which occurs in test instance 7 for all three solving methods—is substantially lower in our solution compared to the manual schedule's and the commercial solution's overtime usage at 101.25, 120, and 159.8 minutes per driver, respectively. Hence, while our solution allocates less average overtime per driver, its primary benefit in terms of overtime is to allocate the least overtime for test instances that require substantial usage of overtime.

Table 2.7. Vehicle capacity for test instances.

Instance	1	2	3	4	5	6	7
Routes driven							
Our approach	13	11	10	10	13	21	27
Manual schedule	17	17	17	13	23	25	31
Commercial solution		13	13	9	16	21	24
Mean available capacity at warehouse departure (excl. compression)							
Our approach	3.2	4.5	2.6	7.3	2.2	1.6	2.6
Manual schedule	5.1	5.1	5.1	13.3	9.2	-0.7	-6.6
Commercial solution		7.8	7.8	7.2	8.4	-3.6	-7.7
Routes with available vehicle capacity ≤ 0							
Our approach	3	1	2	1	3	7	5
Manual schedule	5	5	5	0	7	11	18
Commercial solution		2	2	0	2	13	16
Routes with available vehicle capacity ≤ 3							
Our approach	6	6	7	5	9	18	20
Manual schedule	7	7	7	4	8	12	21
Commercial solution		6	6	2	7	16	20
Routes with available vehicle capacity ≤ 5							
Our approach	11	8	8	6	12	20	24
Manual schedule	7	7	7	4	9	16	23
Commercial solution		6	6	5	7	18	21
Routes with available vehicle capacity ≤ 10							
Our approach	13	10	10	7	13	21	26
Manual schedule	11	11	11	5	12	20	29
Commercial solution		7	7	6	11	21	23

Table 2.7 provides insights into how the vehicles were utilized. Specifically, a contributing reason to the greater cost savings that our solution achieved was that the vehicles were fuller on average when leaving the warehouse for most of the test instances. However, this improved

average does not occur because our solution generally provided full vehicles—the manual schedule and the commercial solution generally provide more routes with no available capacity at the route start. Instead, this improved average occurs primarily because of a steadily good utilization of the vehicles because our solution outperforms the other solving methods in the average filling rate at route start. This is primarily because the manual schedule and commercial solution frequently use far too large vehicles to deliver goods when they do not fill up the vehicles completely, which increases transportation costs and reduces vehicle utilization.

Table 2.8. Waiting time for test instances.

Instance	1	2	3	4	5	6	7
Number of waits at location							
Our approach	7	9	6	7	5	6	5
Manual schedule	4	3	3	4	6	5	9
Commercial solution		2	2	3	14	10	13
Total waiting time at the location (hours)							
Our approach	8.4	4.6	3.1	3.8	3.1	4.3	2.2
Manual schedule	4.4	4.1	4.1	5.9	4.5	3.9	10.4
Commercial solution		0.1	0.1	1.5	14.6	9.5	8.9

Table 2.8 illustrates how waiting times before the time windows affected the total cost savings. The table indicates no clear over- or under-performance of our solution compared to the manual schedule and the commercial solution, with more waits at the location before the time window for four test instances and fewer waits for the remaining three. Similar results can also be observed for the total waiting time at the location.

However, our solution provided a lower variance in waiting time than the manual schedule and the commercial solution. This is a useful metric because it signals how reliable the solving methods are in a critical aspect of vehicle routing planning—namely, planning the routes with respect to time windows. For instance, the greatest under-performance in total waiting time for our solution was 4.0 hours against the manual schedule, whereas the greatest over-performance was 8.2 hours. Similarly, the greatest under-performance against the commercial solution was 4.5 hours, while the greatest over-performance was 11.5 hours. These longer waiting times for the manual schedule and the commercial solution consequently substantially impact the total working day duration and—subsequently—the total driver costs, which drives the total cost savings for the our solution upwards.

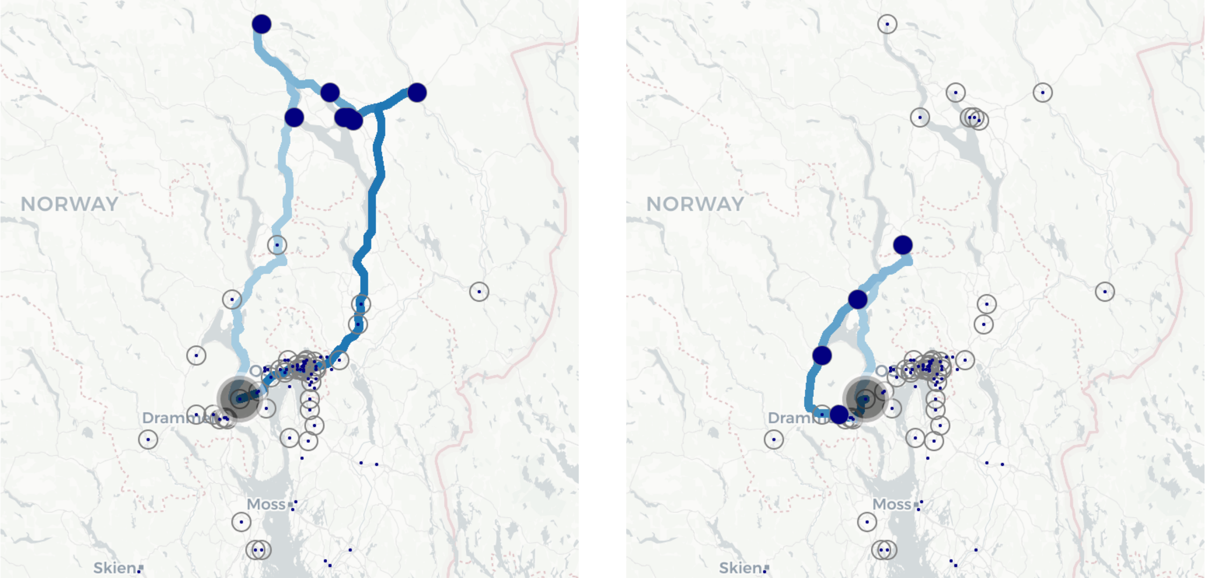

Figure 2.1. Two routes used in test instance 1.

Figure 2.1 shows two of the routes that were selected in test instance 1. The enlarged circle represents the warehouse, the dark circles represents stops on the route, and the outlined circles represents other locations not visited on the route. Furthermore, the edges between these nodes represents the driving directions, color highlighted from a light to a dark shade to represent the start and end of the route, respectively. These maps illustrate well the proximity consideration of the route-generating process. The first map consists of relatively rural locations and so the neighbor set will usually involve just their direct neighbor in two directions if time windows, vehicle capacity, and other considerations allow it. The second map consists of more central locations in the south of the route, and so the southernmost location will have a wider range of locations in its neighbor set. As can be seen in the right-most map, the solver then ended up assigning a trip between two location in the South-West that were not direct neighbors.

2.4.2 Limiting Overtime Usage to the Manual Schedule

Our approach provided a cost-superior solution in Section 2.4.1, but at the expense of higher overtime use for the longest-working driver than in the manual schedule. From a practical perspective, drivers are already used to the current overtime hours set from the manual schedule. However, they may not be interested in a longer maximum working day duration than they currently experience. This implicit precedent is, therefore, an important factor to consider if a company is to change its scheduling practices. Hence, it is interesting to identify the impact on total costs if the optimization model did not allow more overtime usage than the maximum overtime usage of a driver in the manual schedule, which was 3.8 hours for test instance 1. Additionally, it becomes clearer what other factors besides longer working days contribute to the savings.

Table 2.9. Comparing problem variants B and C.

Problem variant	B	C
Minimization objective	Total costs	Total costs
Max working time constraint	15 hours	Max manual work
Total day costs (NOK)	41 556	43 215
Total driving costs (NOK)	12 223	14 703
Total salary costs (NOK)	27 962	27 045
Total toll station costs (NOK)	1371	1466
Total km	1655.8	1941.5
Total km excl. from/to warehouse	497.1	445.9
Drivers used	11	12
Total work time (hours)	79.6	82.6
Total overtime (hours)	6.3	3.4
Average hourly wage rate (NOK)	351.4	327.5
Routes driven	13	13
Mean available capacity at warehouse departure (excl. compression)	3.2	5.4
Number of waits at location	7	7
Total waiting time at location (hours)	8.5	7.6
Variables	120 541	120 541
Constraints	195	195
Relative MIP gap (percent)	2.4	1.7
Running time (seconds)	462.5	464.1

Table 2.9 indicates an overtime usage of 3.4 hours and that the total day costs are only marginally higher in problem variant C (43 215 NOK) compared to problem variant B (41 556 NOK), even though the longest working days have been reduced substantially. This increased cost in problem variant C is primarily due to the increase in total driving costs, which amounts to 14 703 NOK in problem variant C against 12 223 NOK in problem variant B. This disparity can be attributed to the increased total kilometers driven, which sees an increase from 1655.8 km in problem variant B to 1941.5 km in problem variant C.

The increased driving distance of 17.3%, and 29.1% in kilometers driven from/to the warehouse, occurs because the reduced allowable overtime usage per driver enforces multiple routes far away from the warehouse to be allocated. However, the total overtime usage is already low in problem variant B, meaning the benefit from reducing overtime usage is relatively low. Hence, decision-makers must thoroughly analyze overtime needs and avoid setting arbitrary limits per each driver's maximum levels.

Moreover, salary costs are largely unaffected, with a slight decrease in problem variant C. This decrease of 917 NOK occurs due to a reduced hourly wage rate—caused in part by reducing overtime use by 46%—by 6.8%, but is partially offset by an increase in working hours of 3.8%. While the total waiting time at locations decreased in problem variant C, the increased total driving distance of 17.2% contributed to longer working days overall. However, by and large, imposing overtime restrictions increased the total day costs by 4.0%, but this constrained alternative still outperforms the total day costs of the manual schedule by 21.8%.

2.4.3 Minimize Necessary Overtime

While Section 2.4.2 explored the consequences of limiting the working day to the longest working day of the manual schedule, it is also interesting to limit the working day duration but void of the manual schedule's interference. Specifically, to allow the minimum amount of overtime necessary to visit all locations but to disallow a working time longer than that. Furthermore, exploring its consequences on daily costs is interesting if the number of drivers was minimized while still imposing a constraint of working as little overtime as necessary.

Hence, we implemented a pre-problem with the objective function (2.23)—used in problem variant A—that was introduced in Section 2.3.3 to identify the minimum overtime needed. This objective function value was then implemented as the right-hand side in constraint (2.11), which prevented any driver from working longer than the minimum overtime needed to deliver and pick up all goods.

Table 2.10. Comparing problem variants B, D, and E.

Problem variant	B	D	E
Minimization objective	Total costs	Total costs	Drivers used
Apply minimax working time	No	Yes	Yes
Total day costs (NOK)	41 556	45 565	49 199
Total driving costs (NOK)	12 223	17 510	20 476
Total salary costs (NOK)	27 962	26 190	26 049
Total toll station costs (NOK)	1371	1865	2674
Total km	1656	2285	2445
Total km excl. to/from warehouse	497	654.9	519.0
Drivers used	11	13	12
Total work time (hours)	79.6	90.3	86.1
Total overtime (hours)	6.3	1.4	1.1
Average hourly wage rate (NOK)	351.4	290.2	302.5
Routes driven	13	14	13
Mean available capacity at warehouse departure (excl. compression)	3.2	9.7	7.5
Number of waits at location	7	7	4
Total waiting time at location (hours)	8.5	9.8	3.7
Variables	120 541	120 541	120 555
Constraints	195	195	222
Relative MIP gap (percent)	2.4	1.0	0.0
Running time (seconds)	463	60	41

Table 2.10 indicates cost increases when implementing the working day constraint, albeit at a level substantially below the manual schedule's solution. This cost increase of 9.6% and 18.4% for problem variant D and E, respectively, is primarily caused by deliveries at locations far away from the headquarters that takes substantial time driving to, causing a surge in the total distance because more vehicles must be employed to deliver goods to these locations. Consequently, the total driving costs drive the total cost increase for both problem variant D and E.

However, the total salary costs are slightly reduced for these constrained problem variants compared to problem variant B. This decrease of 6.3% and 6.8% for problem variant D and E, respectively, is caused by a reduced average hourly wage rate—which in part occurs because of lower usage of overtime hours. The total working hours, however, increase for these two problem variants, offsetting the cost saving of cheaper labor per hour.

2.4.4 Minimizing Distance

Thus far, the cost-minimization approach proved successful compared to the manual schedule. However, minimizing costs involves many variables, such as each driver's hourly wage rate, overtime, vehicle costs, toll station costs, and more. Consequently, minimizing costs considers the social constraints at least partially, whereas a distance-minimizing problem variant does not—despite its popularity in the literature. Hence, it is interesting to analyze and understand how such a change in the objective would modify drivers' working hours, total day costs, and distance driven.

Table 2.11. Comparing problem variants B and F.

Problem variant	B	F
Minimization objective	Total cost	Distance driven
Apply minimax working time	No	No
Total day costs (NOK)	41 556	57 714
Total driving costs (NOK)	12 223	15 057
Total salary costs (NOK)	27 962	41 158
Total toll station costs (NOK)	1371	1499
Total km	1656	1599
Total km excl. to/from warehouse	497	579
Drivers used	11	8
Total work time (hours)	79.6	91.2
Total overtime (hours)	6.3	37.3
Average hourly wage rate (NOK)	351.4	451.3
Routes driven	13	10
Mean available capacity at warehouse departure (excl. compression)	3.2	2.4
Number of waits at location	7	7
Total waiting time at the location (hours)	8.5	18.4
Variables	120 541	120 542
Constraints	195	196
Relative MIP gap (percent)	2.4	0.8
Running time (seconds)	462.5	7.4

Table 2.11 compares the cost-minimizing problem variant B and the distance-minimizing F. Minimizing distance provided a distance decrease of only 3.4%, whereas the total day costs increased by 38.9%. This cost increase can be largely attributed to the increase in salary costs, which is driven by the 492% increase in overtime and 116% increase in waiting time at the

location. However, since the distance-minimizing objective employs 27% fewer drivers, the overtime usage worsens, with an increase from an average of 0.6 hours to 5.3 hours per allocated driver. Ultimately, a distance-minimizing objective yields few benefits compared to a cost-minimizing objective while substantially increasing the impracticality and costs of its solution, while problem variant B is the more efficient solution in all regards. From the perspective of overtime, a distance-minimizing problem variant proves detrimental because one of the easiest ways to reduce the distance driven is to have each vehicle drive longer—ultimately increasing overtime usage.

2.4.5 Key takeaways

In summary, from our results we derive the following insights:

- Across test instances, our approach outperforms the alternative methods in terms of total day costs, achieving cost savings of 17.4%–36.4% against the manual schedule and 9.7%–25.5% against the commercial alternative. These cost savings are caused mainly by assigning fewer routes and reducing the average distance between each location within a route when excluding the driving to/from the warehouse.

- While our solutions plan for overtime to create a more efficient solution in all test instances, the overtime usage largely outperforms that of the manual schedule and the commercial solution's usage when there are no practical limits to overtime usage.

- Limiting overtime to the smallest allowable overtime usage to undertake all deliveries leads to longer total working hours and longer total driving distance because the most valuable usage of overtime occurs for locations far away from the headquarters when these relatively remote locations are constrained by working day duration more than vehicle capacity.

- A more flexible policy for overtime usage—but within the allowable working limits—positively affects the total day costs and the total working hours of the drivers. While increased overtime fatigues the drivers, all else equal, the reduced total amount of working hours ensures that the drivers can work less on other days during the working week because the same amount of goods can be delivered with fewer resources.

2.5 Concluding Remarks

This article uses a mathematical programming model to assess the impact of social constraints on the vehicle routing problem. The model is implemented to study a single-depot, capacitated, multi-trip model for a heterogenous fleet of trucks with time windows, deliveries, split pickups, asymmetric distances, and real-life speed limits. The model has been applied to real-world instances of a florist retail chain in Norway.

This study shows that cost savings and social constraints are not necessarily trade-offs. By emphasizing and incorporating social constraints both in the mathematical model and in the route-generating process, substantial cost savings were achieved across all test instances. Specifically, the model reduces total costs by 17.4%–36.4% and 9.7%–25.5% against manually produced schedules and a commercial solver, respectively, both attempting to minimize total costs.

Three main findings directly impact working time and salary costs. First, our approach generally assigns fewer routes to deliver the same amount of carts to each location, which reduces total working hours—all else equal. Second, cost savings were generally substantial in our solution due to reduced salary costs. Across all test instances, the total work time was lower in our solution compared to the manually produced schedule and the commercial solution. Third, our solution also selected the most cost-efficient drivers who—including overtime payments—contribute to generally reducing the average hourly wage rate for the utilized drivers.

Furthermore, by studying multiple problem variants in more detail for one of our test instances, we found that limiting overtime use to the manual schedule's longest working day increased the total day costs by 4.0%, but this constrained alternative still outperforms the total day costs of the manual schedule by 21.8%. Hence, allowing more extended overtime use than the manual schedule used did not explain all of the cost savings. Moreover, distance minimization had detrimental impact on total costs when the wages were included in the calculations of the total costs, implying the assignment of more expensive drivers, increase of overtime usage, and increase of waiting time. The effect was particularly adverse for overtime usage, increasing from an average of 0.6 hours to 5.3 hours per allocated driver. Hence, distance minimization might become an unfavorable objective when incorporating these social constraints. We also found that total salary costs decreased when allowing the minimum amount of overtime necessary to visit all locations but disallowing a working time longer than that. However, total day costs increased substantially due to more drivers having to drive longer routes to keep overtime usage as low as necessary. When constraining the overtime usage, total overtime usage decreased from 6.3 hours for 11 drivers to 1.1 hours for 12 drivers while total cost savings relative to the manual schedule reduced from 24.8% to 10.9%. Consequently, the greatest cost savings for the operations of the florist chain may be found by employing a more flexible approach to overtime usage that still respects working laws. While this is a more fatiguing intra-day alternative for drivers, it can be offset by fewer working hours on other days because of the reduced total amount of working hours reduces the demand of the drivers.

As the focus of this article was not algorithmic and we rather studied few real-world instances in detail, interesting future work is to test the model and the route-generating process on a greater span of routing problems and a larger experimental setting. Deriving more general results and analytical conditions as a function of the maximum overtime allowed and other parameters is also an interesting avenue for future research.

2.A Generating Routes

This section describes the route-generation process and the considerations made. The central ideas of this process are twofold. First, primarily good-quality routes should be generated while simultaneously providing enough flexibility for the solver to identify cost-effective alternatives for a fleet of vehicles. While good-quality routes are necessary to provide promising results, keeping a relatively small number of routes will decrease the solving time, all else equal. Second, to explain to decision-makers how the model produced its routes, there need to be specific, explainable rules in the route-generation process, which heuristics and metaheuristics such as genetic algorithms, tabu search, and ant/bee colony optimization generally do not provide.

For each of the test instances provided in this article, the florist chain supplied information about the locations to be visited—including the address, time windows, and the number of carts to be delivered—available personnel, all relevant cost parameters, in addition to the manually made schedule to be used as a comparison. In addition, the florist chain had also tested a commercially available route planning solution whose objective is to minimize total costs while respecting constraints like those presented in this article. For the test instances in which the florist chain had tested the commercial solution, they also provided the automatically produced schedule for comparison.

Naturally, only the locations demanding delivery or pickups intraday were considered in the route-generating process, along with the warehouse. Each location was allocated a set of "neighbors"—the only non-warehouse destinations a vehicle can drive to next from the given origin location. This neighbor set is denoted $\mathcal{N}_i$, of which i is the origin location. From such origin-destination pairs, routes were generated under certain conditions. Sections 2.A.1–2.A.5 cover in detail how these neighbors were generated, while Section 2.A.6 describes the route generation further. Furthermore, each of the neighbor-generating processes in Section 2.A.1–2.A.5 is evaluated on "height and capacity constraints". That is, a neighbor will be added if it passes the processes' conditions and if at least one vehicle in the truck fleet meets both vehicle height requirements at both locations and can capacitate the total delivery demand at both locations.

2.A.1 Proximity

The first and most crucial neighbor-generating process is based on the relative closeness between locations and on a hypothesis of utility. Specifically, we hypothesize that the marginal utility of adding an element to the neighbor set diminishes when adding neighbors ordered by distance, while the solving time increases due to more routes being produced—all else equal. Hence, we utilize a concave threshold curve as seen in Figure 2.A.1 to make it gradually more challenging to add a location to the neighbor set the larger the set already is.

Another consideration is the general observation that locations relatively far from others require fewer neighbors. If a location is relatively rural, there is a high likelihood that a vehicle should visit a direct neighbor because the driving-specific costs are high relative to other delivery-related costs. Consequently, these costs may outweigh the benefits of visiting other, non-direct neighbors, such as potentially utilizing the vehicle's capacity. By comparison, more central locations have relatively low driving-specific costs. Hence, the benefits of visiting non-direct neighbors—such as utilizing the vehicle's capacity better—may outweigh the additional driving-specific costs.

$$threshold_i = \frac{\ln\left(1 + \frac{i \cdot \alpha}{|I|-1}\right)}{\beta} \tag{2.27}$$

Assigning fewer neighbors to rural locations and more to central locations is captured by (2.27), which defines a threshold rule for adding the i'th neighbor to the neighbor set.

First, the hyperparameters $\alpha > 0$ and $\beta > 0$ were decided, which defines how limiting the proximity decision rule should be. Second, an asymmetric distance matrix for all locations is produced using data from the Google Maps API. While costs or travel time could have been used instead because they represent a cost-minimization objective better, the distance was used because it is not sensitive to external factors such as traffic. Third, to avoid sensitivity to outliers for rural locations with relatively large distances, the percentile rank of all values in the matrix was calculated, and the percentile rank matrix was then scaled to $[0; 1]$.

In a scaled percentile rank matrix, rural locations cluster towards a value of one for most of its distances. In contrast, central locations cluster closer toward a value of zero. This characteristic is then utilized in the fourth step, in which—for each location—we produced a lagged difference vector and ordered it from the lowest to the highest value. Fifth, each origin location i was evaluated from the nearest destination location to the furthest. If the scaled distance percentile rank was below the threshold value, the destination location was added to the neighbor set $\mathcal{N}_i$. Once the i'th location surpasses the threshold value, the proximity-generating process is halted for this location, and the initial neighbor set has been defined. Consequently, relatively rural locations had smaller neighbor sets than relatively central locations.

Furthermore, two additional hyperparameters were defined to prevent the proximity process from generating an excessive number of routes. First, an upper limit of four neighbors was decided for the test instances presented in Section 2.4, and the primary factors deciding this hyperparameter value are the number of drivers, vehicles, and locations to visit, in addition to allowed solving time, computational resources, and costs of having a potentially higher optimality gap. Second, a lower limit of two delivery points was decided. While a lower limit could surpass two to enforce more routes to be generated, the lower limit should not be set lower than two because relatively rural locations would get uni-directional routes only, potentially resulting in spurious and costly driving schedules. By enforcing at least two neighbors, the solver can then choose, for example, to drive either northbound from a location or southbound to achieve the most cost-efficient route of the two. Moreover, since the florist chain has multiple pickup points, but most locations are delivery points, it is important to ensure good utilization of the vehicles by avoiding too few delivery points in the neighbor set. Hence, the lower limit is set to include a minimum number of delivery points instead of neighbors—which can be either delivery points or pickup points.

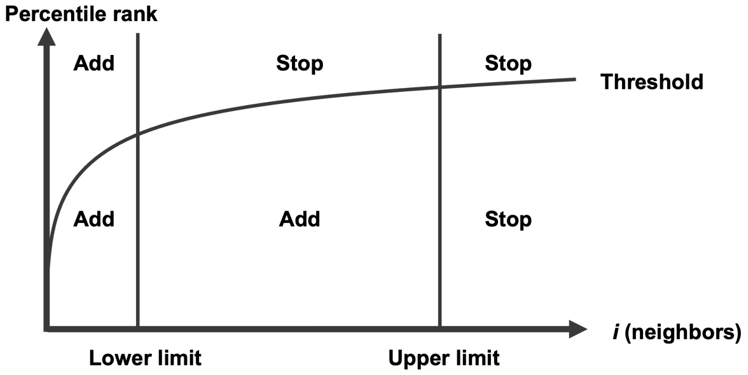

Figure 2.A.1. The neighbor-generating process "proximity". Neighbors are added to the neighbor set as long as the i'th neighbor is below the lower level or both below the upper level and below the percentile rank threshold line.

2.A.2 Overlapping Time windows

A potential challenge when building $\mathcal{N}_i$ using only proximity is that delivery time windows may differ substantially. While it is not a problem if the added driving time still enables the driver to reach both locations with different time windows on one trip with little to no waiting time, it may still be valuable to add additional destination locations to $\mathcal{N}_i$ that overlap substantially on time windows within the working day of the drivers.

Hence, there are three primary considerations when deciding how much to expand the neighbor set due to time windows. First, the less the delivery windows overlap with the working day of the drivers, the greater the need for expanding the neighbor set. This overlap is calculated as the mean overlap between delivery windows of $\mathcal{N}_i$ and the working day. Second, the less the delivery windows overlap, the greater the need for expanding the neighbor set. This overlap is also calculated as the mean overlap, but the overlap is calculated with all pairs of neighbors in $\mathcal{N}_i$. Third, the fewer locations already added to $\mathcal{N}_i$, the more emphasis should be added on expanding $\mathcal{N}_i$ with time-window overlapping neighbors.

The rounded value of equation (2.28) defines how many additional neighbors should be added to $\mathcal{N}_i$. Let θ^w be the total overlap duration between the working day and the locations in $\mathcal{N}_i$, let θ^n be the total overlap duration between all pairs of locations in $\mathcal{N}_i$, and let $\gamma > 0$ and $\delta > 0$ be hyperparameters. The additional number of neighbors to add to $\mathcal{N}_i$ is then:

$$\left(\left(1 - \frac{\theta^w}{|\mathcal{N}_i|}\right) + \left(1 - \frac{\theta^n}{2(|\mathcal{N}_i| - 1)}\right)\right)^{\gamma} \left(\frac{|I| - 1}{|\mathcal{N}_i|}\right)^{\delta} \tag{2.28}$$

Similarly to the proximity rule, there are a minimum and maximum number of new neighbors that can be added to $\mathcal{N}_i$, which for the test instances presented in this article were set to 0 and 2, respectively. Finally, the new neighbors added to $\mathcal{N}_i$ were decided based on proximity to the origin location i, given that they had largely overlapping time windows.

2.A.3 Parallel Pathways (adding more directions)

Thus far, locations have been added to $\mathcal{N}_i$ based mainly on their proximity to the origin location. However, if many of the locations already added to $\mathcal{N}_i$ lie along the same driving path, a driver can practically drive in only two directions—for example, Eastbound or Westbound. Hence, there is little variety in the routes that will be generated that involve the origin location i, and it would be valuable to add more locations to $\mathcal{N}_i$ that increase this variety—say, also having the option to drive Northbound or Southbound, too. Hence, the third decision rule in the neighbor-generating process attempts to remove locations in $\mathcal{N}_i$ that largely share the same properties of driving direction and time windows and replace them with other near-proximity locations that allow the driver to go in a different direction. The decision to remove these locations is intentional as largely overlapping neighbors generally provide little value but substantially increase the number of routes that will be generated.

There are three primary considerations regarding overlapping pathways. First, a neighbor may only be considered for removal from $\mathcal{N}_i$ if it is sufficiently far away from the origin location, which was set to 10 km for the test instances in this article. This minimum value is because locations near the origin will likely share a substantial portion of the driving route despite the destinations being located in different directions. Second, mainly rural origins will have neighbors that share large portions of their driving route, and this replacement process should not be stricter in replacing neighbors for rural origins than for more centrally located origins. To achieve this, we used the percentile ranks of the origin location's percentile rank vector, which gave all origin locations the same linear percentile rank values after ordering by proximity. Third, parallel roads must also be considered. Specifically, if two vehicles drive on parallel roads and go in the same direction, these roads must be treated as the same. For the test instances in this article, any road within 1 km of each other was considered parallel.

The first step in this process is to identify which locations in $\mathcal{N}_i$ to be replaced by others. First, for each pair of neighbors in $\mathcal{N}_i$, the ratio of overlapping driving paths from the origin to each neighbor was calculated and stored in a matrix. Then, for each neighbor in $\mathcal{N}_i$, we identified which neighbors it shared a pre-defined ratio threshold with—set to 70 % for the test instances in this article—and recorded the furthest away location of the two as removal candidates from $\mathcal{N}_i$.

The second step in this process is identifying which locations should be added to $\mathcal{N}_i$ instead. First, we identified the destination locations not in $\mathcal{N}_i$ and not overlapping beyond the ratio threshold with any neighbor in $\mathcal{N}_i$ before ordering the remaining replacement candidate list by the closest to furthest proximity to the origin location. Then, we excluded any replacement candidate whose delivery size was greater than the largest vehicle available that day could contain after deducting the delivery size to the origin location i. That way, we ensured that at least one vehicle could drive both to the origin and destination on a single route. Finally, we iterated through the set of removal candidates, replacing each with the nearest remaining replacement candidate. Consequently, $\mathcal{N}_i$ has more variety in the driving direction from the origin location without increasing its cardinality.

2.A.4 Trajectory

Thus far, each neighbor-generating process has added locations to $\mathcal{N}_i$ primarily based on its proximity to the origin location. While the extensive focus on proximity is natural from a cost- and distance-minimizing perspective, shorter detours between two locations may also contribute positively to these objectives. Hence, the fourth process identifies a pre-defined (set to 1 in the test instances) number of locations not in $\mathcal{N}_i$ that lay the closest to the trajectory between the origin location i and the warehouse. Suppose these locations involve a detour of less than a pre-defined number of minutes of driving time—set to a maximum of 15 minutes additional driving time than driving directly from the origin location i to the warehouse. In that case, these locations should be added to $\mathcal{N}_i$.

The primary benefit of this neighbor-generating process is the increased variety of routes. For example, and as is demonstrated in Figure 2.A.2, the Northern-located Parkgården adds the Southern-located Oslo Central Station to its neighbor set because of the trajectory process. Since Oslo Central Station has a neighbor set containing multiple Southern-located locations such as Gunerius, the routes may consist of multiple deliveries in entirely different regions. This variety of routes is a valuable addition to constructing $\mathcal{N}_i$ because it allows more efficient utilization of vehicles when routes are not confined to the same region, mainly when orders to each store are so great that it is challenging to utilize the truck otherwise.

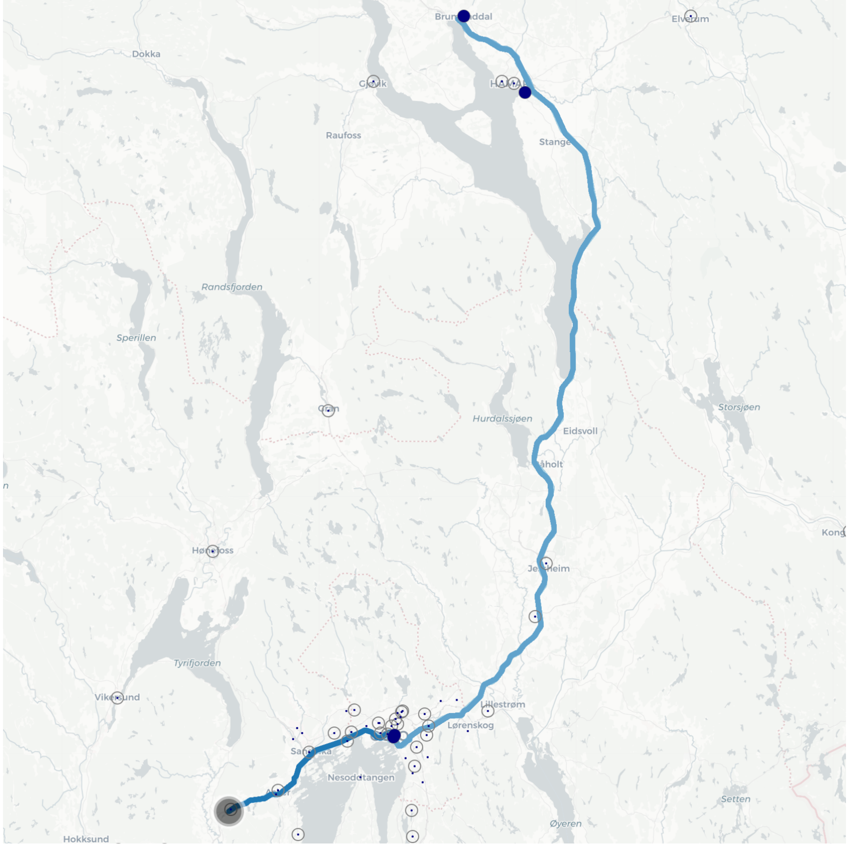

Figure 2.A.2. After delivering goods to Vikasenteret and Parkgården (North), the vehicle was allocated to take a detour to deliver to Oslo Central Station and Gunerius in the city center of Oslo before returning to the warehouse.

2.A.5 Hardcoded

The final neighbor-generating process involves the florist chain's expertise. While the previous neighbor-generating processes create cost-efficient routes compared to the other solving methods, there may be nuances and subtleties in the route generation that the automatic neighbor-generating processes might overlook. If so, it is valuable if the experienced transportation planners can manually define neighbors in the neighbor set, and the hardcoded neighbor-generated process allows just that. This hardcoded process is unnecessary for each test instance but was implemented at the initial development phase in discussions with the florist chain. For the test instances presented in this article, only one connection—between Stortorvet Senter and Re-torvet—was hard-coded into the neighbor set on the days both had deliveries.

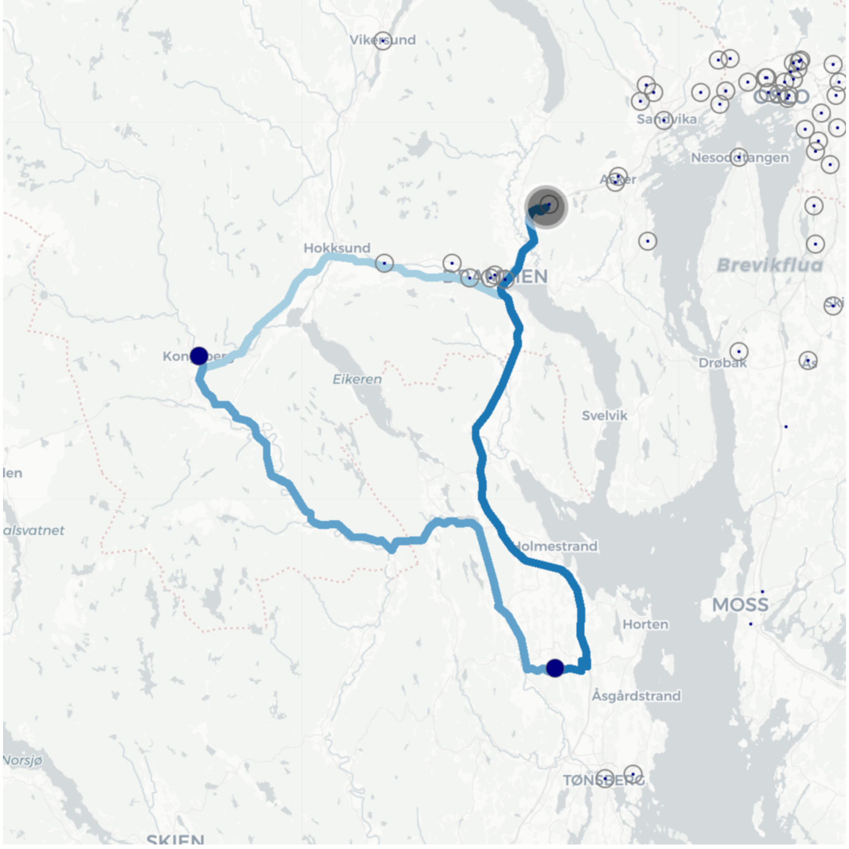

Figure 2.A.3. A vehicle was allocated to visit Stortorvet Senter first and then Re-torvet because of the hardcoded addition

2.A.6 The Route-Generating Process

After the neighbor-generating process, we begin stacking these neighbors together into routes. This process follows two primary stages, which are described next.

The first stage is to expand the route set and consists of two steps. First, we begin an initialization phase, in which a one-stop route is produced for each location that demands deliveries or pickups that day—excluding the warehouse. Second, we expand the set of routes of length j to length $j + 1$. The expansion is performed by appending all locations in the neighbor set of the last location of the route, excluding locations already added to the set. For instance, consider a two-stop route—first to visit Oslo Central Station and then Gunerius. Then, three-stop routes are generated from this route by appending all neighbors in the neighbor set of Gunerius that are not already added to the route. Assuming the neighbor set of Gunerius consists of Oslo Central Station, the National Theatre, and Frogner Atrium, then the newly generated routes will be Oslo Central Station–Gunerius–National Theatre and Oslo Central Station–Gunerius–Frogner Atrium.

The second stage involves pruning the route set for each iteration of j beyond a lower threshold. All routes of length j are grouped by locations they visit, ignoring the order in which they are visited. For each route within these groups that contains at least two routes, we calculate the additional driving time than the other routes within the group. If this percentage exceeds a pre-defined threshold, the route is excluded from the route set. While this pruning process has the advantage of reducing the number of routes, it risks eliminating potentially viable routes that could have been more cost-efficient when considering routes of a greater length than j. Hence, the pre-defined threshold of additional driving time was set to a high value of 50% to reduce this risk.

The route set is expanded until a stopping rule is met. These stopping rules are that 1) the time limit for route generation has been reached (set to thirty minutes), 2) the maximum number of routes to be produced has been reached (set to two million routes), 3) no vehicles can deliver to the existing routes plus any other neighboring locations in the neighbor set of the last stop of the existing routes, or 4) no routes can add more neighboring stops without violating the working day duration when also factor in driving both from the warehouse to the first stop and from the last stop back to the warehouse. When the stopping rule is triggered, the warehouse located is both prepended and appended to the route, completing the route. Figure 2.A.4 shows one of the routes that was generated and ultimately selected as one of the delivery routes that day.

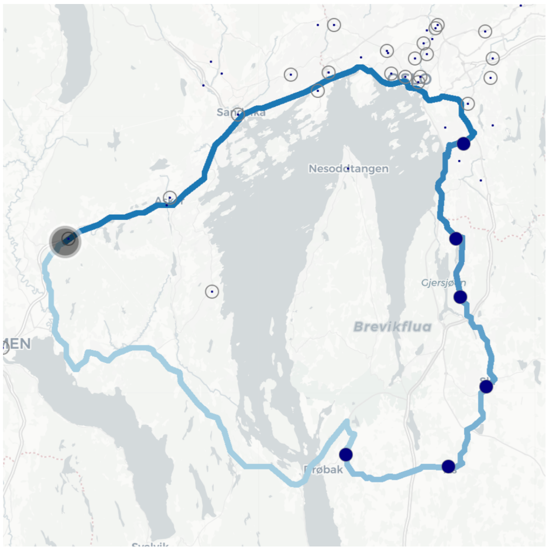

Figure 2.A.4. A generated route.

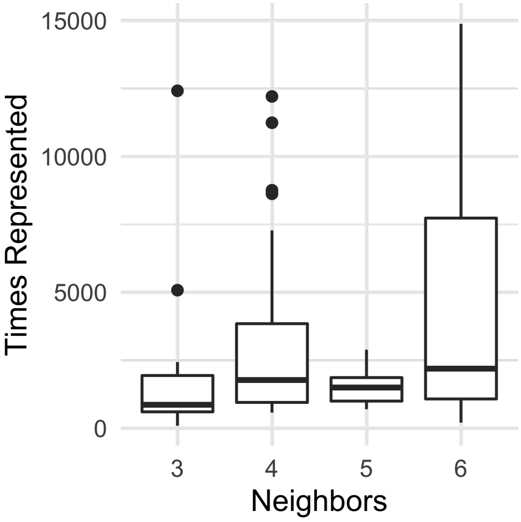

Figure 2.A.5. How many times stores have been represented in routes given their number of neighbors in test instance 1.

Feasibility is a primary concern when generating such routes. Hence, each location must be represented in a wide array of routes to increase the likelihood that a cost-efficient solution can be produced while respecting vehicle capacity, vehicle height limitations, working day constraints, and all other constraints concerning the problem. Figure 2.A.5 illustrates how frequently each store has been represented for a given number of locations in the route of test instance 1, in which the median representation is 1454 times with a minimum of 94 times for a test instance with 25 275 produced routes. This frequent representation of each location consequently reduced the likelihood of infeasibility problems and increased the likelihood of cost-efficient solutions—as demonstrated in Section 2.4.1.

2.A.7 Adding Route Details

After generating the routes, we add route details. First, we create the driving schedule of all activities, predicted starting and ending times, container load changes, total driving time, and distance driven in kilometers for each route. For driving activities, we used the Google Maps API to collect directions, distance, and total driving time for all pairs of the 88 locations with deliveries or pickups on the given day. Second, we extend the route details of each route with driver- and vehicle-specific details, including available capacity, costs, and more. Third, we identify the toll stations passed, the time of passing each station, whether a cost should be incurred or if the 'hour-based rule' is still respected from a previous passing, and if the cost occurs, we identify what the cost level should be given the vehicle driven.

The Google Maps API offered three settings—best guess, pessimistic, and optimistic. Our empirical experience showed that the 'best guess' setting tended to be overly optimistic, often leading to overtime and narrowly missed delivery windows. This experience aligns with the finding of Eglese et al. (2006) and Ehmke et al. (2012) that underestimating travel times is the most common problem transport managers report. As a large portion of the driving occurred during rush hours, when the 'best guess' setting proved to be especially unreliable, we decided to utilize the 'pessimistic' setting. This setting allowed us to plan for potentially longer driving times deliberately, lowering the risk of overtime.

Considering the vehicle fleet's physical characteristics was another crucial factor in our driving cost calculations. Each vehicle in the heterogenous fleet had a specific set of attributes, including the weight when empty, the maximum loading weight, the height, the container capacity volume-wise, the average fuel consumption, the average vehicle cost per kilometer, the engine type, and the eligible drivers. In addition, specific locations had constraints on which vehicles could access their docks, typically due to height or weight limitations.

A further consideration was the time required for unloading at each location. This unloading time was divided into two components: the “fixed” time needed for parking and departure and the “variable” time required per container to be unloaded. This data was available both for each store and for the warehouse. Additionally, the warehouse had different time estimates for loading and unloading the trucks—specifically, 5 minutes to prepare the truck for loading and unloading, with 40 seconds per container for the former and 20 seconds per container for the latter. Five minutes was also added at the end of the route for other administrative tasks.

There was a requirement to maintain additional space in the truck during container pickups to ensure smooth operations. This space was necessary as picking up carts could affect the ability to unload other carts designated for subsequent stops. Specifically, the required space was the greater amount of 15 percent of available capacity or four carts. Therefore, the practical capacity for picking up carts at a given location was the vehicle capacity minus the required space minus the total number of carts yet to be delivered.

Container compression was allowed in the morning if it was financially beneficial. This process involved moving goods onto other carts to reduce the number of carts, with a penalty of 10 minutes per container removed. Compression was only permitted for candidate routes with a load of $\langle 100; 105]$ percent of the vehicle's capacity.

Finally, toll station costs were calculated. First, we used the geographic coordinates of the toll stations and matched them with the Google Maps waypoints. Second, we categorized each toll station according to its zones and the chargeable direction(s). In the case of one-way toll stations, we automatically evaluated the direction of the Google Maps waypoints to identify whether the route passed through the payment direction. Third, we identified the costs for each vehicle in the fleet— regarding vehicle weight and engine type and for each day in question's hours. Fourth, the 60-minute free pass rule was evaluated. Specifically, if the vehicle were predicted to pass Indre ring and Oslo-ringen multiple times or pass Bygrensen multiple times within 60 minutes, then the toll station would not charge the vehicle. Using the route details, we predicted the time the vehicle would pass each toll station on the route and evaluated whether this hourly rule was respected. If it was not, the cost of passing the toll station in question was added to the total route cost.

Autonomous Ferries in Light of Labor Regulations—A Passenger Perspective

Christian Braathen, Julio C. Goez, and Mario Guajardo

Abstract

We study the potential effects of introducing autonomous ferries in a transportation system of water buses. We develop two integer linear programming models and a heuristic to find weekly transportation plans both for a fleet of autonomous ferries and manually operated ferries, of which the objective is to minimize the total penalty occurring from the number of passengers, arrival delay penalties, and alternative transportation penalties. The models differ because working laws affect the crews' working capacities, and we study the changes when these requirements are absent with autonomous ferries. Our work is motivated by the case of Bergen, a coastal city in Norway. In this case, the use of autonomous ferries has the potential to improve passengers' utility significantly. Our results suggest that it may be beneficial to consider autonomous ferries as a complementary alternative that can operate especially in low-demand hours—a recommendation that may be particularly relevant if there are few autonomous ferries available or the ferries can only be operative for a limited number of hours of the day.

Keywords: public transportation; assignment problem; autonomous ferries; integer linear programming

JEL Classification: **J22** Time Allocation and Labor Supply; **R41** Transportation Demand; **C61** Optimization Techniques; **C44** Operations Research

3.1 Introduction

The use of autonomous vessels (AVs) is a highly topical issue due to the increasing capabilities of artificial intelligence and changing regulations allowing their use in national and international waters. The Finnish archipelago was a significant milestone in the development of AVs with the implementation of the world's first fully autonomous ferry, the "Falco," in 2018 (Rolls-Royce Holdings plc, 2018). This implementation serves as a proof of concept for the potential use of AVs in public transportation. However, to this date, too little focus has been given to the end-users of autonomous vessels. While the development of AVs has focused heavily on the technology and the literature has focused primarily on the feasibility, implementation, and assessment of autonomous ships, there has yet to be a good understanding of the potential gains of end-users from using autonomous vessels compared to manually operated vessels.

In this article, we deepen the understanding of AVs' potential by exploring how passenger demand is met if ferries are autonomously operated instead of manually. We develop two integer linear programming models and a heuristic to find weekly transportation plans for manually operated and autonomous fleets of ferries. The models differ because working laws affect the crews' working capacities, and we study the changes when these requirements are absent with AVs. Hence, given the same fleet size and passenger demand, we explore whether autonomous ferries meet this demand better than manually operated ferries—and if so, to which extent.

The models produce a one-week schedule for autonomous and manually operated ferries and are applied to a case in Bergen, Norway. Each model seeks to minimize passenger delay within its constraints, including the constraints related to labor laws and crew to different extents. While the approach of this article is to take a public perspective and understand differences in passenger delays, the approach can easily be extended to a ferry operator's perspective to understand the differences in revenue potential.

The article is structured as follows. First, Section 3.2 explores the existing literature. Section 3.3 presents the problem together with an integer linear programming model, while Section 3.4 develops a heuristic approach to solve it. A computational study and results are presented in Section 3.5. Section 3.6 concludes with some final remarks.

3.2 Literature

A substantial part of the literature on autonomous ships—including ferries—centers around the feasibility, implementation, and assessment of autonomous ships. The literature has focused extensively on Norway as the country leads the technological development of autonomous ships (Munim, 2019). This literature is developing, and several open questions still need to be answered before autonomous ships can be safely used at scale in real-life settings (Gu et al., 2021). Thus far, the current literature has focused on technological advancements, safety considerations, and legal and economic concerns.

3.2.1 Technological Advancements

Most studies on autonomous ships focus on the technological perspective (Munim, 2019), ranging from a project-wide focus to specific problems such as maneuverability, collision avoidance, and docking. Kurt and Aymelek (2022) define the level of autonomy in ships across six levels, from low automation to completely crewless ships, of which most projects to date have yet to achieve the highest level of autonomy.

One project achieving the highest level of autonomy, however, is the “Falco” car- and passenger vessel in the Finnish archipelago—the world’s first fully autonomous ferry and began its operations in 2018 (Rolls-Royce Holdings plc, 2018). This project was a milestone in advancing autonomous ferries as most previous attempts at autonomous ships have involved remotely, but still manually, operating the ships. Hence, the project in Finland demonstrated the feasibility of the project considered in Bergen, Norway.

3.2.2 Safety Considerations

Beyond the technological advancements of autonomous systems, a key reason autonomous shipping has gained tremendous momentum in recent years is safety concerns (Xu et al., 2023). While it is believed that human error will not disappear entirely with autonomous ships, 70–90% of marine accidents today are believed to be caused by such human errors (Porathe et al., 2018). Hence, there is a substantial potential to reap from a safety perspective by transitioning to fully autonomous vessels, and the literature has been heavily focused on this perspective (Gu et al., 2021).

3.2.3 Legal Concerns

When assessing manually operated vessels against autonomous vessels, one essential difference between the two treatments is the working laws protecting the crews. The protection and working limitations of each worker differ because they may be set by multiple parties such as national law and labor unions. However, there are three essential rules to protect most workers—defining how long each work week can be, the minimum amount of rest that must take place between shifts, and the minimum amount of consecutive rest that must take place each week. By Norwegian working laws, a work week can typically be 40 hours long, there must be at least 11 hours of rest between each shift, and each worker must have at least 35 hours of consecutive rest at least once per week (Ministry of Labour and Social Inclusion, 2005).

There are two common approaches to scheduling the weekly rest. A frequently used, albeit simplified, alternative to scheduling these resting days is to schedule five-day work weeks and two consecutive days off. By comparison, a more elaborate approach is to define which shifts each worker should be assigned to and then identify the workers' resting times given the abovementioned requirements. (Van den Bergh et al., 2013)

Most personnel scheduling literature covered by the literature review of Van den Bergh et al. (2013) has the following characteristics. First, it usually studies full-time contracts. Second, it studies individuals instead of whole crews. However, if assuming that all individuals are homogeneous and each shift needs just the same amount of crew members, then each crew can be considered as one individual. Third, the shift sequence is typically decided by a scheduling model, in contrast to the assignment of tasks, groups, or time. Fourth, shifts are typically allowed to overlap. These four characteristics are also followed in this article. Moreover, a slight majority of the personnel scheduling literature has a fixed shift length, while a slight minority has a definable one. The heuristic in this article has the former characteristic, while the exact models has the latter.

3.2.4 Economical and Operational Concerns

Several studies attempt to improve aspects of the economics and operations of autonomous vessels. These studies typically involve different perspectives on operating costs and capital costs. For instance, Aslaksen et al. (2021) study the effects of combining fixed schedules and on-demand services provided by a fleet of autonomous ferries in Germany. Employing both an operational cost and a capital cost perspective, Gu and Wallace (2021) address a problem involving decisions on facility location, fleet allocation, and routing in the operations of a water-taxi case in Norway. Employing a capital cost perspective, Reddy et al. (2019) make the case that the alternatives to zero-emission, autonomous passenger ferries are bridges and manned ferries and that the former is more environmentally friendly, flexible, and cost-effective. Besides passengers, the literature on cargo transportation has also discussed economical and operational aspects of autonomous vessels. For instance, employing an operational perspective, Zhang and Wang (2020) attempt to minimize fuel cost consumption and delay penalties of autonomous vessels by assigning berths at each port and deciding the vessel's arrival time at each port. Akbar et al. (2021) schedule autonomous vessels to identify the optimal fleet of vessels—in terms of number and size—in addition to the route of each vessel from Europe to Norway and vice versa. Also, Nguyen et al. (2022) investigate economic and operational effects of introducing autonomous vessels to liner shipping networks. Their findings suggest that autonomy can render savings, especially due to lower time charter costs and lower bunker costs.

Moreover, a frequent economic argument for using autonomous ships is that *personnel costs* will be reduced. This argument is discredited by Kooij and Hekkenberg (2021), who study how autonomous systems affect crew sizes and argue that autonomous systems will reduce the need for various crew members for *phases* of the trip—such as the sailing, arrival and departure, and loading and unloading phases—and not the entire trip itself. Until full autonomy is achieved, the authors demonstrate that introducing autonomous systems may *increase* the total costs as autonomous systems face operating and maintenance costs. Since ships that have not yet reached full autonomy will not necessarily be able to reduce the number of staff as each role could be demanded for at least one phase of the sailing, the total costs may increase.

Even if full autonomy is achieved, total costs—excluding cost savings from fewer human-made marine accidents—may still not be reduced. Specifically, Kretschmann et al. (2017) studied the cost benefits of autonomous vessels of an autonomous vessel project. The authors found that the reduced crew wages alone are likely not enough to offset the increased costs of fully autonomous vessels—such as increased shore control center costs, increased demand for maintenance crews, and increased capital costs for autonomous ship technology.

While the previous studies on reducing the number of personnel do not indicate a reduction of total costs from reducing crews, even with fully autonomous systems, little has been said on whether fully autonomous systems have the potential to improve *revenues*. Specifically, the crews' working times are limited by working laws, whereas fully autonomous systems are not. In a practical setting, this could imply that fully autonomous systems, such as autonomous ferries, could operate at all hours of the day except for when maintenance is needed. By comparison, manually operated ferries would also be limited by the number of crews available to work within the constraints of the working laws, which could cause a poorer utilization of the fleet of ferries and ultimately lead to fewer passengers being transported than autonomous ferries. Hence, there may be an increased revenue potential from using a fully autonomous fleet.

The revenue potential can be identified by studying changes in the volume of passengers. Specifically, as revenue depends on both pricing and quantity, and deciding prices is a separate decision that affects passenger demand, it is valuable to isolate the problem by assuming fixed prices for autonomous and manually operated ferries. Hence, the revenue problem is reduced to a passenger quantity question, which we attempt to explore in this article by studying how much better autonomous ferries may meet the passenger demand throughout a whole week compared to manually operated vessels. As far as we know, this research question has yet to be explored in the literature.

3.3 Integer Linear Programming Formulation

With the changing regulations allowing fully autonomous vessels in national and international waters, understanding the impact on passengers of using autonomous ferries instead of manually operated ferries is highly topical. When studying this problem, it is necessary to form constraints regarding the ferries, crews, and passengers working on the manually operated ferries available.

Additionally, to add depth to our analysis, we introduce the concept of a "passenger group"—a group of passengers of various sizes sharing the exact origin, destination, and preferred departure time. The purpose of passenger groups is to create a realistic variation in the demand for ferries throughout the week-long schedule while reducing the complexity of solving the problem, as whole passenger groups can be allocated instead of each passenger. Since the ferries are assumed to have no capacity constraints, the models will either allocate a ferry to transport all or none of the passengers in the passenger group.

For ferries, it is important to ensure that they are defined logically, each being maximally at one location at a time and arriving from the location it arrived at in the previous time slot. The ferries must either be operative or on stand-by at a stop, of which the latter is defined as allocating the ferry to a two-stop route of one time slot's duration, with the origin and the destination being the exact location. These routes in which a ferry is on stand-by are called "trivial routes", and these routes are beneficial if waiting for a future passenger group generates a higher objective function value than starting another route or transporting a passenger group at the present time. Additionally, for manually operated crews, there is a limitation on the number of ferries that can be used simultaneously if the number of crews available is less than the number of ferries available.

Crews face three constraints. Specifically, crews cannot work longer than eight-hour workdays, must rest for at least 11 hours after each shift, and must have at least one 35-hour block of continuous rest each week.

Passenger groups will either use ferries as their mode of transportation or use alternative transportation. The passengers can only be assigned to a ferry if a relevant route has been assigned during the time window in which passenger groups are willing to wait for the ferry. If no such ferry has been allocated during this time window, the passenger group will use alternative transportation—such as a bus.

We formulate two models with the same objective function to compare schedules with crewed and autonomous ferries. Since the ferries are supposed to be a solution to offload to the existing transportation network, two considerations must be made in the objective function in both models. First, to attempt to transport as many passengers as possible who prefer a ferry, while the second consideration is to minimize the arrival delay the transported passengers face. By employing the same objective function and a subset of constraints on the same test instances, we can isolate the effect autonomous ferries have compared to the manually operated ones.

When exploring to which extent autonomous ferries may meet passenger demand better than manually operated ferries, the following assumptions are made about ferries and crews. First, we assume the ferries are homogeneous, with the same transportation time for autonomous and manually operated ferries, and without capacity limitations or maintenance needs. Second, we assume homogeneous crews and that the same crew members always work together so that we can treat them as one joint entity of workers. Additionally, we assume that no overtime is allowed for all crews, shift changes between crews are instantaneous, and the ferry does not need to reach a particular stop to perform the shift change.

Assumptions are also made for schedules, docks, and passengers. First, we assume no fixed patterns are necessary for the ferry schedule, such as line 1 running every 30 minutes daily during daytime hours. The transportation time between two stops is deterministic and equally long for manually and autonomously operated ferries. Second, we assume unlimited docking capacity, meaning there will be no congestion when multiple ferries simultaneously arrive at the same stop. Third, we assume that no passengers want to start and end at the same stop, that they will travel with one ferry only to reach their end destination, and that they are willing to be delayed to their arrival destination for a limited amount of time when using the ferry—otherwise, they will use alternative transportation methods such as a bus.

Table 3.3.1. Model components.

Sets	
T	A circular set of time slots.
R	Set of routes.
V	Set of ferries in the autonomous operations model and the set of crews in the manual operations model.
P	Set of passenger groups.
S	Set of ferry stops.
Subsets	
$\mathcal{P}_p \subset (R, T)$	Set of route–departure time alternatives for passenger group p.
$\mathcal{R}_t \subset (R, T)$	Set of ongoing routes at time t.
$\mathcal{V} \subset R$	Set of trivial, vacant routes staying still on a particular stop for one timeslot.
$\mathcal{S}_{st} \subset (R, T)$	Set of routes starting at stop s at time t.
$\mathcal{E}_{st} \subset (R, T)$	Set of routes ending at stop s at time t.
$\mathcal{O} \subset (P, R, T)$	Tuple set of (passenger, route, departure time) options.
$\mathcal{H}_t^{11} \subset T$	Set of 11-hour contiguous time slots starting at time t.
$\mathcal{H}_t^{19} \subset T$	Set of 19-hour contiguous time slots starting at time t.
$\mathcal{H}_t^{35} \subset T$	Set of 35-hour contiguous time slots starting at time t.
Parameters	
$\delta_{prt} \geq 0$	The arrival delay that passenger group p faces by taking route r that begins at time t.
$\rho_p \geq 0$	Number of passengers in passenger group p.
$\alpha_p > \max\limits_{p,r,t} \delta_{prt}$	Penalty if passenger group p must use alternative transportation.
$\nu \geq 0$	Max number of ferries operating simultaneously.
$\eta \geq 0$	Max timeslot per week the ferry/crew can work.
$\mu_v \geq 0$	Min timeslots per week that ferry/crew v must work.
$\theta_r \geq 0$	Duration of route r, as measured in time slots.
Variables	
$x_{vrt} \in \{0, 1\}$	Ferry v starting route r on time t.
$y_{prt} \in \{0, 1\}$	Passenger group p is assigned to the route r that starts its route at time t.
$y_p^A \in \{0, 1\}$	Passenger group p is assigned to alternative transportion.
$z_{vt}^R \in \{0, 1\}$	Ferry/crew v rests on time t.
$z_{vt}^D \in \{0, 1\}$	Ferry/crew begins its daily rest on time t.
$z_{vt}^W \in \{0, 1\}$	Ferry/crew begins its weekly rest on time t.

3.3.1 Passenger Constraints

$$y_p^A + \sum_{\substack{(r,t) \in \mathcal{P}_p: \\ (p,r,t) \in \mathcal{O}}} y_{prt} = 1 \quad \forall p \in P \tag{3.1}$$

$$y_{prt} \leq \sum_{v \in V} x_{vrt} \quad \forall (p, r, t) \in \mathcal{O} \tag{3.2}$$

Constraint (3.1) states that each passenger group either goes on a specific route or uses an alternative transportation method like a bus. Alternative transportation must be allocated if the passenger group is not picked up during the *departure window*, consisting of the preferred departure time and the three time slots ahead.

Moreover, passengers can only be assigned if a ferry arrives for relevant routes and time slots. This condition is captured by constraint (3.2). In this context, the relevant routes are the routes that contain both the origin and destination locations of the passenger group, with the origin being visited before the destination on the route. The relevant time slots are the starting time for the route that makes the ferry arrive at the origin location within the passenger group's departure window.

3.3.2 Working Requirements Constraints

Trivial routes have one primary function for autonomous ferries and two functions for manually operated ferries. For both types of operations, it may be beneficial for the objective function if a ferry does not transport passengers at a particular time slot but instead waits at a specific location before departing at a later time slot—for example, because a large number of passengers would like to depart only a couple of time slots later. Hence, it is important not only to capture allocations of ferries in use but also their geographical position when they are not in use so that they depart from this position next. By assigning the ferry a series of trivial routes when the ferry is not operational, this benefit can be captured.

$$\sum_{r \in \mathcal{V}} x_{vrt} + \frac{1}{11} \sum_{\tau \in \mathcal{H}^{11}_{prev(t,10)}} z^{D}_{v\tau} \geq \left(1 + \frac{1}{11}\right) z^{R}_{vt} \quad \forall v \in V, t \in T \tag{3.3}$$

Additionally, manually operated ferries use trivial routes to define whether a particular crew is resting, a critical concern when employing working requirements to the model. Hence, we define z^{R}_{vt} to equal 1 when a particular crew and their ferry v rests—i.e., not working—on time t. This rest is defined to occur when two criteria are met. First, the crew's ferry is assigned a trivial route at time t, meaning no passengers are transported, and the ferry is on standby. Second, the crew is undertaking their daily rest of minimally 11 hours at time t. Constraint (3.3) defines when z^R_{vt} is equal to zero.

$$\sum_{r\in R}\sum_{t\in T}\theta_r x_{vrt} - \sum_{t\in T} z^R_{vt} \leq \eta \quad \forall v \in V \tag{3.4}$$

$$\sum_{r\in R}\sum_{t\in T}\theta_r x_{vrt} - \sum_{t\in T} z^R_{vt} \geq \mu_v \quad \forall v \in V \tag{3.5}$$

Furthermore, the x variable denotes a ferry's location and route. While it does not denote if the ferry is actively being used, we do know the hours the crew rests. Hence, we can deduct those hours to get the actual number of hours worked to enforce that each crew can maximally work a $\eta = 40$ hour work-week, which constraint (3.4) states. By a similar approach, we can enforce a minimum work-week duration per crew of μ_v, which constraint (3.5) states.

$$\sum_{\tau\in\mathcal{H}^{11}_t}\ \sum_{\substack{(r,\dagger)\in\mathcal{R}_\tau:\\ r\notin\mathcal{V}}} x_{vr\dagger} \leq |\mathcal{H}^{11}_t|(1 - z^D_{vt}) \quad \forall v \in V, t \in T \tag{3.6}$$

$$\sum_{\tau\in\mathcal{H}^{19}_t} z^D_{v\tau} \geq 1 \quad \forall v \in V, t \in T \tag{3.7}$$

$$\sum_{\tau\in\mathcal{H}^{35}_t}\ \sum_{\substack{(r,\dagger)\in\mathcal{R}_\tau:\\ r\notin\mathcal{V}}} x_{vr\dagger} \leq |\mathcal{H}^{35}_t|(1 - z^W_{vt}) \quad \forall v \in V, t \in T \tag{3.8}$$

$$\sum_{t\in T} z^W_{vt} \geq 1 \quad \forall v \in V \tag{3.9}$$

Moreover, a crucial element of working requirement constraints is the constraints concerned with resting times. First, we must enforce that the daily rest is at least 11 hours long and how often they should minimally occur. When enforcing the former, it is sufficient to define when the daily rest does not occur—which constraint (3.6) states is the case if no ongoing routes have been assigned to crew v on time t. Additionally, we must ensure that the daily rest occurs minimally every $11 + 8 = 19$ hours since workdays cannot be longer than 8 hours long, and constraint (3.7) ensures this. Second, we must enforce that at least one 35-hour period of continuous rest occurs at least once during the week, which requires two constraints. Similarly to the daily rest, it is sufficient to define when the weekly rest does not occur, and constraint (3.8) follows a similar approach for the weekly rest as constraint (3.6) does for the daily rest. Finally, constraint (3.9) ensures that a minimum of one such 35-hour resting period must be enforced throughout the week.

3.3.3 Ferry Constraints

$$\sum_{(r,\tau)\in\mathcal{R}_t} x_{vr\tau} \leq 1 \quad \forall v \in V, t \in T \tag{3.10}$$

$$\sum_{(r,\tau)\in\mathcal{E}_{s,prev(t)}} x_{vr\tau} = \sum_{(r,\tau)\in\mathcal{S}_{s,t}} x_{vr\tau} \quad \forall v \in V, s \in S, t \in T \tag{3.11}$$

$$\sum_{v\in V} \sum_{(r,\tau)\in\mathcal{R}_t} x_{vr\tau} \leq \nu \quad \forall t \in T \tag{3.12}$$

Beyond the passenger and working requirements constraints, three more constraints are needed to target the ferry logic specifically. That is, a ferry can maximally be at one route on each time slot, and ferries can only begin a route at a stop they arrived at in the previous time slot. Constraints (3.10) and (3.11) ensure these conditions, respectively.

Moreover, constraint (3.12) states that a maximum capacity of simultaneously operating ferries must be respected. While this constraint is inherently respected by the set V in the autonomous model, this constraint serves as a constraint on the maximum number of crews we employ simultaneously if the number of crews and ferries differ.

3.3.4 Variable Constraints

Finally, each variable has an integer constraint.

$$x_{vrt} \in \{0, 1\} \quad \forall v \in V, r \in R, t \in T \tag{3.13}$$

$$y_{prt} \in \{0, 1\} \quad \forall p \in P, r \in R, t \in T \tag{3.14}$$

$$y_p^A \in \{0, 1\} \quad \forall p \in P \tag{3.15}$$

$$z_{vt}^R \in \{0, 1\} \quad \forall v \in V, t \in T \tag{3.16}$$

$$z_{vt}^D \in \{0, 1\} \quad \forall v \in V, t \in T \tag{3.17}$$

$$z_{vt}^W \in \{0, 1\} \quad \forall v \in V, t \in T \tag{3.18}$$

3.3.5 Objective Function

The objective function is to minimize the total penalty occurring from the number of passengers ρ in passenger group p, the delay penalty relative to the earliest possible arrival δ_{prt} for passenger group p, and the penalty α_p occurring because the passenger group chose an alternative transportation method.

$$\min_{y,y^A} \sum_{p \in P} \rho_p \left(\sum_{(r,t) \in \mathcal{P}_p} \delta_{prt} y_{prt} + \alpha_p y_p^A \right) \tag{3.19}$$

3.4 Heuristic

While the integer linear programming formulation covers the entire problem under the assumptions mentioned earlier, a test instance with many passenger groups, in particular, may render the model computationally intractable. Specifically, the biggest driver for the number of decision variables in a test instance is the number of passenger groups, which depends on the number of departure stops, arrival stops, the number of time slots in the schedule, and how many of these (departure stop, arrival stop, time slot) permutations have a non-zero demand. The duration of a time slot should preferably be short—such as five minutes—and there should preferably be a demand for transportation in many of the time slots to create realistic test instances. Moreover, the entire schedule should be at least one week long when considering labor regulations to capture the most important constraints, such as daily and weekly rest. With a week-long schedule consisting of short time slots with a non-zero demand, the number of passenger groups will be high, eventually making the exact model time-consuming or even computationally intractable.

Consequently, a heuristic method to address the problem is valuable as it balances the trade-off between solution quality and computational efficiency. While adhering to the same assumptions as the exact model, our heuristic introduces an additional assumption regarding crew scheduling—that crews are assumed to work eight-hour shifts, commencing simultaneously each day for five consecutive days. This simplification will likely result in a lower solution quality than the exact model for realistic test instance sizes. However, it remains a reasonable assumption in light of the Norwegian Working Environment Act, which emphasizes consistent working hours and mandates a continuous weekly rest period for employees.

3.4.1 Simplified Example of the Heuristic

Before describing the technical details of the heuristic with Algorithm 1 on page 83, it is valuable first to present a simplified example of the approach the heuristic attempts to achieve when allocating a passenger group. For clarity, the term "departure window" is used throughout the example and refers to the intersecting set of time slots between the shift's duration and the time slots of earliest departure to latest departure. The term “arrival window” is also frequently used and has a similar definition, but is instead applied to the earliest and latest *arrival* times.

The first two steps involve the destination only. First, it is evaluated if there are any open time slots during the arrival window to allocate the destination. If no slots are available, the allocation process for this iteration ends. Second, we evaluate if the origin location is already allocated to the departure window. If true, we attempt to allocate the destination only. However, if false, we attempt to allocate the origin and destination locations to the schedule.

The third step involves both the origin and the destination. If the second step is true, we evaluate if we need—and then attempt—to allocate both the origin and destination or if it is sufficient with just the destination. Moreover, when attempting to assign the origin or the destination at a specific time slot, the algorithm identifies if it is feasible considering the already allocated locations in the schedule.

To explain the main steps of the algorithm using a simplified example, assume we are considering the shift $S := \{10, 11, ..., 17\}$ for a manually operated ferry, and we have the five following passenger groups that want to depart from their origin and arrive at their destination during the shift S. The first four passenger groups will be allocated to the schedule, but the fifth is not.

Table 3.4.1. Passenger group details.

Passenger group	A	B	C	D	E
Objective function value if no delay	20	15	10	10	5
From	1	2	3	3	3
To	2	4	4	5	1
Earliest departure	10	11	10	10	10
Latest departure	13	14	13	13	13
Earliest arrival	11	13	11	11	12
Latest arrival	14	16	14	14	15

Table 3.4.2. How the schedule is being produced for passenger groups A–E.

	Time							
	10	11	12	13	14	15	16	17
A	**1**	**2**						
B	1	**2**		**4**				
C	1	2	**3**	**4**				
D	1	2	**3**	4	**5**			
E	1	2	3	4	5			

Allocating passenger group A

Since the schedule is empty and the sets of departure and arrival windows are non-empty, passenger group A is allocated. Specifically, the arrival window has open time slots, so the first step succeeded. Then, it is identified that the origin location still needs to be assigned in the departure window. Consequently, the algorithm must try to allocate both the origin and the destination to the schedule. The algorithm attempts to allocate the origin and destination as early in its departure and arrival window as possible, respectively. Since the schedule is empty before the allocation and the earliest departure and arrival times are contained in the shift, we allocate passenger group A to leave from location 1 at time 10 and arrive at location 2 at time 11. Hence, no arrival delay occurs, and the highest objective function value for passenger group A has been achieved.

Allocating passenger group B

Passenger group B gets allocated at its preferred departure and arrival, too, but leaves a restricted gap in the schedule. Specifically, the algorithm first confirms that the arrival window has open time slots. Then, the algorithm confirms that the origin location has already been assigned to the departure window before evaluating if allocating only the destination to the schedule is sufficient.

At least one of two conditions is met to allocate only the destination. First, if the origin location already occurs in the departure window at least once and there are gaps in the schedule after the first occurrence. Second, if the origin location already occurs in the departure window at least once and the destination location already occurs after the first origin occurrence. Since there are gaps in the schedule after time slot 11—in which the only origin location is found in the departure window—it is indeed sufficient to allocate only the destination, which the algorithm attempts to do next.

The algorithm does this by considering the existing schedule and seeing if the destination can be allocated during the arrival window when traveling from the last allocated location since departure to the destination. Since no locations have been added after the origin at time 11 yet, the destination location can be allocated at the earliest arrival time—time slot 13. Hence, no arrival delay occurs, and the highest objective function value for passenger group B has also been achieved.

Allocating passenger group C

Passenger group C also gets allocated, but not at its preferred departure and arrival, and the schedule only updates the origin. First, the algorithm confirms that there are open time slots in the arrival window of time slots 11–14, with gaps occurring at times 12 and 14. Second, the algorithm identifies that the origin location 3 still needs to be assigned to the departure window. Hence, it proceeds with its attempt to allocate both the origin and the destination to the schedule. At this step, the algorithm first checks if the schedule has already been populated with a journey between the origin and location at the preferred times. If true, we can allocate the passenger group without modifying the schedule.

The next step is identifying the origin and when the origin can be allocated. To identify this, we must find the first gap in the schedule in which it is feasible to depart to the origin from the last location before this gap. Similarly, we must find if it is feasible to depart from the origin at the gap time and reach the next location. At this step, the algorithm sequentially iterates through the time gaps in the departure window—which are potential departure times from the origin—τ^D until a successful allocation has been made. There is only one gap in this window, which is time 12. The previous allocation prior to time 12 is location 2 at time 11, and the subsequent allocation is location 4 at time 13. Since the shortest travel time from location 2 to the origin location 3 is only one time slot, and since the shortest travel time from the origin location 3 to the already scheduled destination location 4 is only one time slot, allocating the origin location to $\tau^D = 12$ is feasible.

Still inside the iteration of $\tau^D = 12$, a similar albeit slightly different approach is followed to identify when the destination location can be allocated. First, the algorithm sequentially iterates through assignable arrival time slots. These slots are empty gaps within the arrival interval and the time slots in which the destination has already been allocated. This set is filtered further to keep only the times after or on τ^D plus the shortest travel time between the origin and the destination. For passenger group C, this set consists of time slots 13 and 14 as one time slot is needed to depart from the origin at $\tau^D = 12$ to the destination. The algorithm begins iterating with $\tau^A = 13$, which is a feasible time slot because it had already been allocated the destination location from a previous passenger group—in this example's case, by passenger group B.

The final step for passenger group C is to allocate the origin but not the destination. When the algorithm identifies a (τ^D, τ^A) pair that is feasible, it ends the process for the passenger group in question and assigns the origin and destination to this time slot, respectively. While passenger group C's destination location 4 was already allocated to time slot 13, the schedule is updated with the group's origin location 3 to depart at time slot 12.

Allocating passenger group D

Similarly to passenger group C, group D also gets allocated but not at its preferred departure and arrival. However, contrary to passenger group C—which was assigned an origin location because the destination was already allocated—passenger group 4 must be allocated both the origin and the destination. However, it follows the same steps as passenger group C, except that the first τ^A is an empty time slot, and the schedule must be updated by allocating location 5 to time slot 14.

Allocating passenger group E

Passenger group E, however, does not get scheduled because the latest acceptable arrival in this example is three time slots after the earliest arrival at time 15. Specifically, the ferry arrives at location 3 after the earliest departure time but must go through locations 4 and 5 first. When done at location 5 at time 14, the ferry spends two time slots to get to location 1 at time 16. This time slot is later than the latest acceptable arrival time, so passenger group E will not be allocated to this schedule. Hence, this ferry transports passenger group A–D—the four passenger groups with the highest individual objective function values—while passenger group E uses alternative transportation.

3.4.2 The Heuristic Formulation

In this section, we describe the core components of our proposed solution approach, which consists of an algorithm to determine the work schedule for a crew and ferry. The primary algorithm—Algorithm 1—is responsible for constructing the work schedule by evaluating multiple alternatives. The main intention of this algorithm is to allocate passenger groups to fit gaps in the schedule so that the number of allocated passenger groups can be high and their delay and alternative transportation penalties can be low.

There are also supporting functions to this algorithm described in the appendix. First, Function 1 is embedded within Algorithm 1 and is responsible for allocating only the destination location of a passenger group since the origin location is already allocated within the passenger group's departure window. Second, Function 2 is also embedded within Algorithm 1 and will attempt to allocate both the origin and destination location inside of the given passenger group's departure and arrival window, respectively, while meeting other constraints.

Choice of algorithmic approach

The heuristic follows a greedy approach, which was selected for five reasons. First, it is easily explainable. Second, a greedy approach can find a feasible, well-performing solution for larger test instances with a high passenger group-to-vessels ratio, such as our computational study involving 10 000 passenger groups and up to 15 vessels to be scheduled over 2016 time slots. This concern is particularly relevant for the manually operated model variant, as its complexity is greater than the autonomous vessel problem variant. Third, while subsequent articles may explore more advanced heuristics, the primary focus of this article is to study the effects of autonomous and manually operated ferries. By avoiding heuristics depending on random seeds, we remove one variable that affects our analysis while simultaneously providing a good benchmark for which future articles can employ. Fourth, each passenger group has a narrow departure window of only three time slots, with few stops. Hence, the number of candidate solutions that would outperform a local optimum that the greedy approach produces is likely low and of marginal improvement. Fifth, the heuristic generated five schedules with deterministic starting points to reduce the disadvantage of starting with a potentially poor starting point.

Regarding the second reason, a primary concern about the greedy approach is that an allocation is not updated later. While this concern is present in all test instance sizes, it is particularly relevant when the number of passenger groups is low—such as 200—and when planning manually operated ferries. Specifically, an underlying assumption of the heuristic is that each crew is allocated to work an eight-hour shift starting at the same time for five days straight, whereas the exact model does not have such a strict requirement. If the number of passenger groups is low, there will be fewer passenger groups per time unit demanding a ferry than if the number of passenger groups is high. Hence, it may be more beneficial to use varying starting and ending times of each shift—as long as the model's constraints are still respected. One example could be for a crew to work from 8 a.m. to 4 p.m. on Monday and 9 a.m. to 5 p.m. on Tuesday because the crew is able to transport one more passenger group by doing so—which is percentage-wise a far greater number than if transporting an additional passenger group when the total number of groups is high. Since the number of passenger groups in most test instances the heuristic is applied to in this article is 10 000, this concern does not affect the total results too much.

The heuristic

The algorithm aims to create a work schedule efficiently for each ferry in a fleet. The algorithm is designed to iteratively evaluate potential workweeks, starting with initializing an empty `WorkSchedule` list that stores the most favorable schedule for each ferry, considering the various constraints and objectives.

For each *ferry*, the algorithm initializes an empty `PotentialWorkWeek` list and assesses various *schedule starting times*—the first day only for the autonomous ferries and each day of the week for crews. This list represents the best-chosen schedule for a given starting block and allows us to compare and select the most suitable workweek allocation.

It then examines each *working day* within a week—seven days ahead of the week for autonomous ferries from the schedule starting time, and five days ahead for crews. Due to the circular nature of the work schedule, the weekdays are treated as a circular set, and the algorithm uses this circular set to generate the time slots that occur in this potential shift.

The *starting row* loop is introduced to account for the fact that, although passenger groups are sorted by their objective function values when scheduled to reach their destination at their preferred arrival times, the overall objective function value of the entire schedule may be lower if the (time, origin, destination) permutation of the highest-ranked passenger group turns out to be an unfavorable starting point for the schedule at large. By exploring different starting points, the heuristic aims to enhance the overall objective function value, resulting in a better schedule with more passenger groups being transported or facing less delay from preferred arrival times. In our experiments, it became clear that any starting rows beyond five did not provide added value.

Furthermore, the algorithm filters passenger groups based on departure and arrival windows. It initializes an empty `local_schedule` list to construct and store feasible work schedules throughout the algorithm's execution and an empty `passenger_groups_allocated` vector to keep track of the passenger groups that have been allocated to ferries in this local schedule, providing a means to assess the overall efficiency and effectiveness of the alternative work schedules that are produced. The algorithm then attempts to allocate unassigned passenger groups to available time slots, considering whether the origin or both the origin and destination need to be allocated. The algorithm employs the supporting functions that handle the allocation process when allocating passenger groups.

After evaluating and comparing the working day schedules for different starting rows, the algorithm selects the best working day schedule and adds it to the list of potential workweeks. Finally, it identifies the best potential workweek for each crew and ferry and accordingly updates the list of allocated passenger groups.

Algorithm 1: Allocate passenger groups to ferries

```
Definitions:
    S                  set of time slots for the working shift.
    ed                 earliest time slot the passenger group can depart from the origin.
    ld                 latest time slot the passenger group can depart from the origin.
    ea                 earliest time slot the passenger group can arrive at the destination.
    la                 latest time slot the passenger group can arrive at the destination.
    schedule start     day 1 for autonomous ferries and day 1,2,...,7 for manually operated ferries.
Input: list of ferries, list of schedule starts, data frame of passenger group data, list of time slots
Output: finalized work schedule for each ferry
initialize an empty WorkSchedule;
for each ferry do
    initialize an empty PotentialWorkWeek;
    for each schedule start do
        for each working day from the schedule start do
            initialize an empty WorkingDay;
            generate time slots for potential shift;
            for starting row 1..5 of the passenger groups data frame do
                move rows < starting row to the end;
                filter to passenger groups in which $ld \in S \vee ea \in S$;
                initialize an empty local_schedule and a local copy of allocated passenger groups;
                for each passenger groups not allocated to local_schedule or previous ferries yet do
                    if any open time slots in $\{ea, .., la\} \cap S$ then
                        if the origin location is already assigned to local_schedule in $\{ed, .., ld\} \cap S$
                          then
                            if it is sufficient to allocate destination only; then
                                attempt to allocate destination only to local_schedule (Function 1);
                            else
                                attempt to allocate both origin and destination to local_schedule
                                  (Function 2);
                            end
                        else
                            attempt to allocate both origin and destination to local_schedule
                              (Function 2);
                        end
                    end
                end
                evaluate local_schedule and update the WorkingDay list;
            end
            find the best-performing local schedules across different starting rows;
        end
        add the best-performing local schedule to the list of potential workweek schedules;
    end
    find the best potential working week schedule for the ferry, which is the final schedule for the ferry;
    update the list of already allocated passenger groups;
end
```

3.5 Computational Study

Similar to many urban areas, Bergen, Norway faces transportation challenges. With a growing population, a surge in tourism, and public transportation that resembles a hub-and-spoke structure—with most buses going via the city center due to the coastline and mountainous surroundings—the city's transportation system is congested and under pressure to accommodate the increasing demand for mobility.

This challenge spurred the interest of the "blue light rail" project (Opus, n.d.). The project aims to provide an alternative mode of transportation for the city's residents and tourists to reduce traffic congestion and improve the overall transportation experience. The concept of the project is still in the planning phase, and further research is needed to evaluate its feasibility. However, despite being in the planning phase, the project has gained significant interest because the fjords are an integral part of the city of Bergen and play a significant role in shaping its history, culture, economy, and political identity. The current suggestion involves six stops for the blue light rail, and Table 3.5.1 shows the hypothetical travel times between each of these stops, as measured in time slots used in the models presented in this article.

Table 3.5.1. Travel time between each stop, measured in time slots.

	1	2	3	4	5	6
1	0	1	2	3	2	1
2	1	0	1	2	3	2
3	2	1	0	1	2	2
4	3	2	1	0	1	2
5	2	3	2	1	0	1
6	1	2	2	2	1	0

The popularity of a passenger group was calculated as follows. First, each station, weekday, and hour are given popularity scores, with weekends and nights being notably less popular than other weekdays and hours. Second, static popularity scores of the origin and destination locations, weekday, and hour are multiplied. A popularity parameter is then calculated as $pop_{pg} := \ln(1 + \prod_{i=1}^{4} p_{i,pg})\ \forall pg$, in which p is a vector that represents each of these four metrics and pg represents the passenger group. Third, the popularity score determines the number of passengers in a passenger group, which is calculated as $\left\lceil (pop_{pg} - \min_{j} pop_j)^{(1/\beta)} \right\rceil \forall pg, \beta > 0.$

Fourth, only n of these passenger groups are kept for the test instance, in which the probability of being drawn is $\frac{pop_{pg}}{\sum_j pop_j}$ $\forall pg$.

Realistic routes were created and added to the set of routes, which is an input to the models. When creating this set, all two-stop routes were generated, including the six "stand-by routes"—one for each stop. Moreover, the majority of the 1200 permutations for routes of length three to five stops are unrealistic because of substantially larger distances traveled than alternative routes involving the same stops. Hence, only 100 of these were kept. For routes of length six, each stop is the origin location for two routes—one going clockwise and the other going counter-clockwise, visiting all stops. Routes of length seven are the same as the routes of length six, except that the origin location is appended as the destination location. The number of stops per route and their frequencies in the set of routes is shown in Table 3.5.2.

Table 3.5.2. The number of routes per given length.

Stops	2	3	4	5	6	7
Frequency	36	76	12	12	12	12

3.5.1 Smaller Test Instances

Computations were undertaken on a Linux Ubuntu 18.04 computer with an 18-core, hyper-threaded 2.6 GHz Intel Xeon Platinum 8272CL processor and 144 GB RAM. The exact models' test instances were run with the mathematical programming language AMPL and the commercial solver CPLEX version 12.10.0—of which CPLEX ran each instance for six hours unless optimality was found earlier—and the heuristic was run in a parallelized mode of R version 3.6.3.

Table 3.5.3. Results with the exact models.

Instance	1		2		3	
Passenger groups	200		200		10 000	
Time slots	168		2016		2016	
Ferries	1		15		15	
Crews	1		15		15	
Objective function value, no assignments	10 554		10 512		524 448	
Treatment	A	M	A	M	A	M
Heuristic						
Objective function value	6730	9016	517	694	108 587	234 658
Exact model						
Objective function value	5553	8333	0	NA	235 642	NA
Relative MIP gap (percent)	0	0	0	NA	100	NA
Variables	43 241	43 745	4 854 850	4 945 568	5 658 150	5 748 868
Constraints	17 537	18 380	228 128	351 149	1 031 430	1 154 449
Running time (seconds)	2.4	9263	2839	NA	21 600	NA

Table 3.5.3 highlights the challenges of the exact models. Specifically, test instance one is a small problem consisting of 200 passenger groups to be assigned, one-hour long time slots, and one ferry and crew. This test instance solves to optimality for autonomous and manually operated ferries, but the latter problem variant takes over 2.5 hours to solve to optimality for a small problem.

Test instance two further highlights the challenges for this manually operated test variant. While the number of passenger groups is still 200, the number of time slots has increased to five-minute time slots, and both the number of ferries and crews has increased to 15. This change to a more realistic setting increases the number of decision variables 113-fold and the number of constraints 19-fold. At this problem size, CPLEX cannot run the problem variant. By comparison, the autonomously operated test variant can still solve the problem to optimality, but the solving time increased from 2.4 seconds to 47 minutes between the two test instances.

However, the autonomously operated problem variant faces challenges when the number of passenger groups increases from 200 to 10 000. While 6003 passenger groups are allocated to ferries, the MIP gap is still 100%, and the objective function value is dramatically improved in the heuristic. Hence, the exact models produce poor results at realistically large problem sizes.

3.5.2 Larger test instances

While the exact models are inefficient for larger test instances, the heuristic allows this possibility. Hence, the following section studies 300 test instances with 5-minute time slots and 10 000 passenger groups demanding transportation. Such small time slots and such a high number of passenger groups result in a more realistic representation of the real-life demand of ferries, and these time slot sizes and passenger group numbers result in an average of 4.96 passenger groups per time slot demanding transportation.

Description of test instances

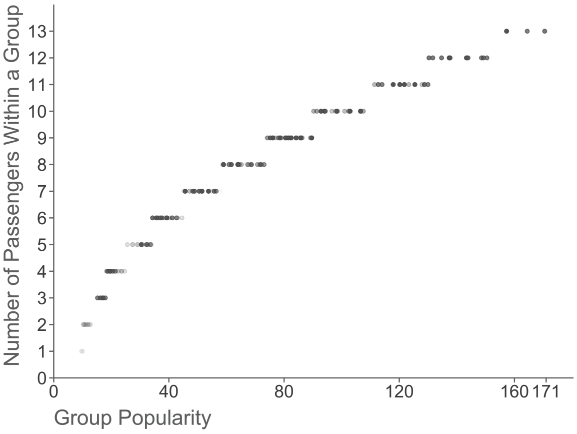

Figure 3.5.1. Number of passengers given the passenger group's popularity.

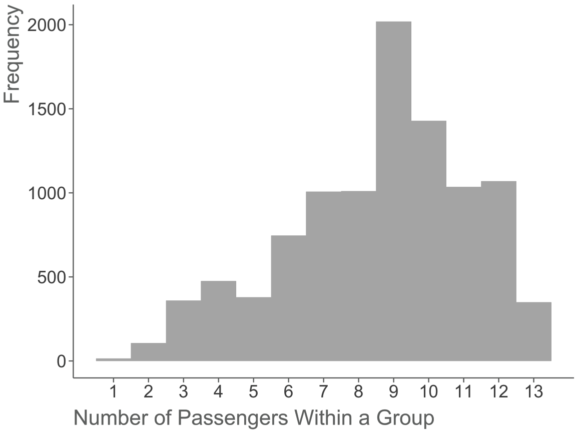

Figure 3.5.2. Distribution of passenger group sizes.

Figures 3.5.1 and 3.5.2 show the distribution of passengers and groups. With six origins, five destinations, and five-minute time slots, there are 60 480 (origin, destination, departure time) permutations during a week. 10 000 passenger groups are then drawn and given a non-zero demand, both of which are drawn based on the group's "popularity"—which is the demand of leaving a particular origin at a specific time slot to go to a particular destination. The number of passengers within these groups is illustrated for test instance 1 in Figure 3.5.1, with a mean of 8.6 passengers per passenger group and a standard deviation of 2.6 passengers. The most frequent passenger group size consists of 9 passengers, and the majority of the passenger groups have 9–12 passengers. Throughout the week, 85 920 passengers will be transported in test instance 1 if all passenger groups are allocated.

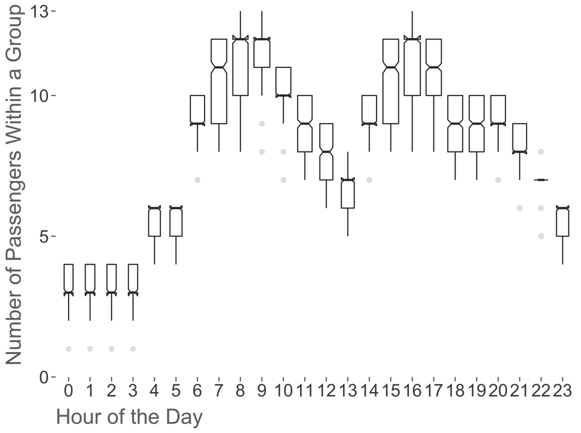

Figure 3.5.3. Distribution of passenger group sizes given preferred departure hour.

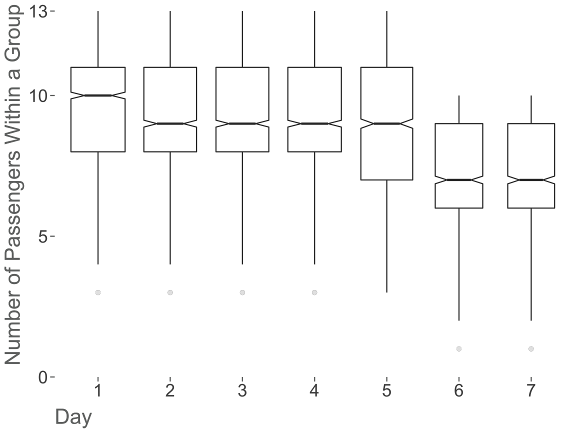

Figure 3.5.4. Distribution of passenger group sizes given preferred departure day.

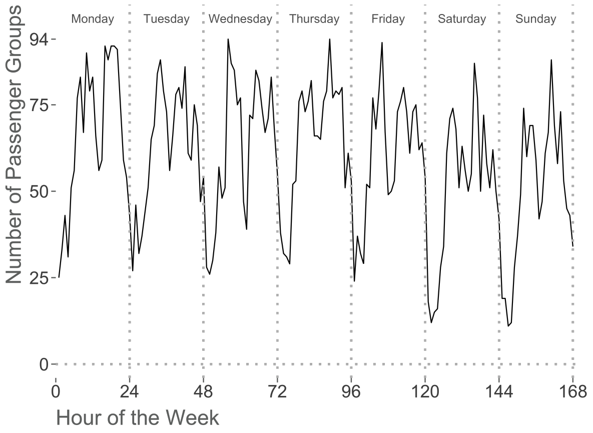

Figure 3.5.5. Number of passenger groups per hour of the week.

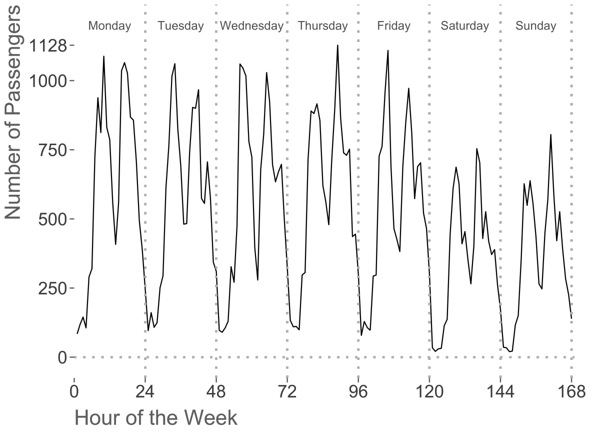

Figure 3.5.6. Number of passengers per hour of the week.

The number of passenger groups and passengers also varies throughout the day and week. Specifically, Figures 3.5.3 and 3.5.4 show that the number of passengers within each passenger group is higher Monday–Friday and in the typical rush hours. Additionally, the number of passenger groups fluctuates during the day and week as well, which Figure 3.5.5 illustrates. Consequently, as the number of passengers is the sum product of the number of passenger groups and the number of passengers within each group, the number of passengers demanding transportation throughout the week fluctuates with rush hours and weekdays—as can be seen in Figure 3.5.6.

Description of solution

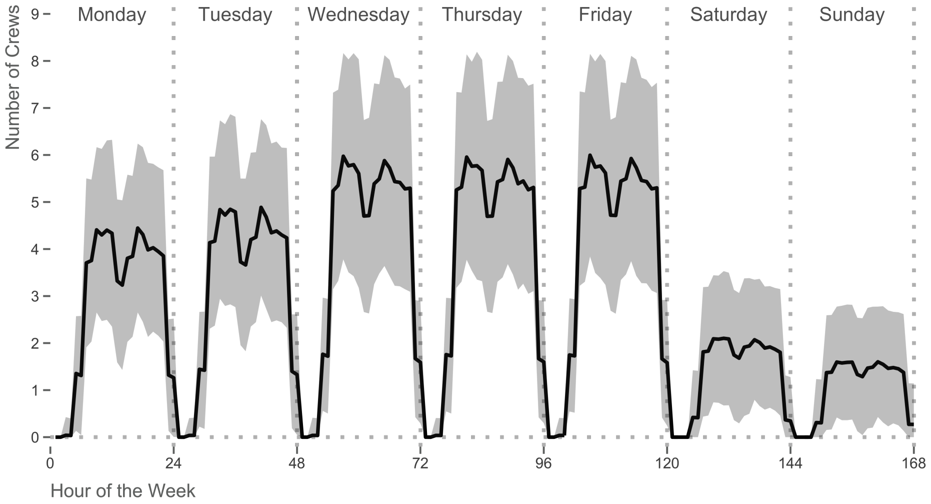

Figure 3.5.7. The highest number of crews working simultaneously for each hour of the week.

Since the models' objective is to minimize the objective function penalty, it is natural that crews are assigned in high numbers during the day and week at peak hours. This effect is demonstrated by Figure 3.5.7, in which none of the 15 crews are scheduled for the first hours of each day, and most crews are allocated for the peak hours during the day—with the most crews being allocated in the morning and afternoon.

Consequently, many simultaneously working crews create an extensive fleet of ferries used only a few hours during the day. For instance, when 15 crews were working, the highest number of simultaneously working crews across all test instances was 8. These under-utilization tendencies exist across all 300 test instances. Hence, the following calculation was made for each test instance to fairly compare autonomously and manually operated ferries for a given fleet size. First, for each crew 1 through 15, we identified how many crews worked on each time slot. Second, we identified the highest number of simultaneously working crews for each crew, equivalent to the lowest number of ferries needed for the schedule. This lowest number of manually operated ferries is the number we compare against an equally sized fleet of autonomous ferries within each test instance.

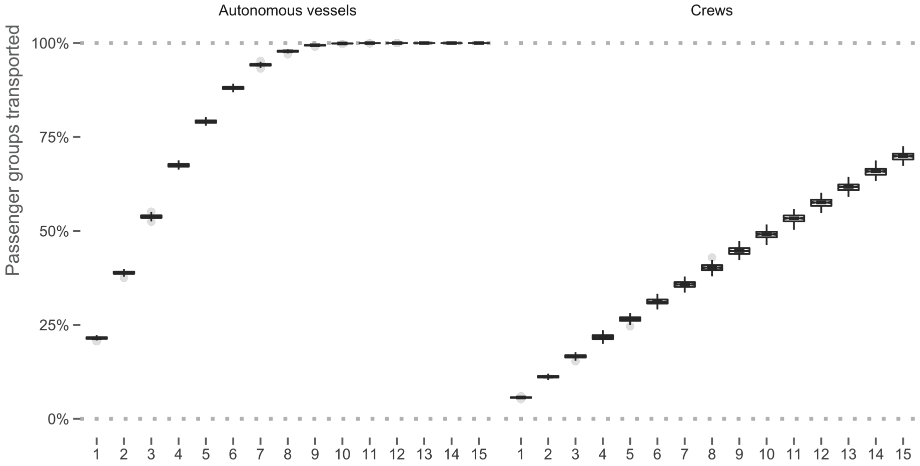

Figure 3.5.8. Passenger demand met by ferry fleet size, as measured in the ratio of passenger groups being allocated.

The low fleet utilization naturally manifests itself in meeting less demand, too. Figure 3.5.8 shows a near-linear relationship between the number of crews transporting passengers and the ratio of passengers being transported, indicating that even with a two-digit number of crews, there are still substantial gaps in the schedule with plenty of demand for ferry transportation. In comparison, autonomous ferries reach a median coverage of 100% across test instances with 12 autonomous ferries and 97.8% with eight ferries. While the percentage difference does decrease between the two types of ferries as the autonomous ferries converge towards 100% coverage, autonomous ferries still cover minimally 43.1% more passenger groups than manually operated ferries for any number of ferries/crews below 15.

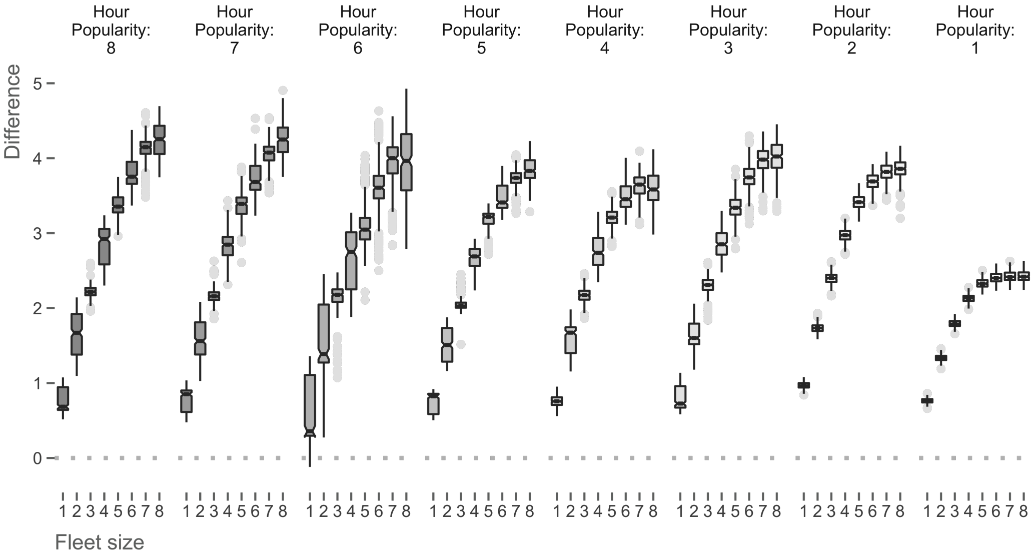

Figure 3.5.9. Surplus coverage by autonomous ferries per time slot compared to manually operated ferries.

Each hour of the day was assigned one out of eight ordinal scores of how many passengers could be expected to demand a ferry ride, all else equal. Generally, the score is lower in the night and higher during the day, with the lowest scores at 00:00–03:59 and the highest scores at hours 08:00–09:59 and 16:00–16:59. As can be seen from Figure 3.5.9, autonomous ferries transport more passengers per time slot across all hour popularities than manually operated ferries. This difference increases the greater the hour popularity and fleet size.

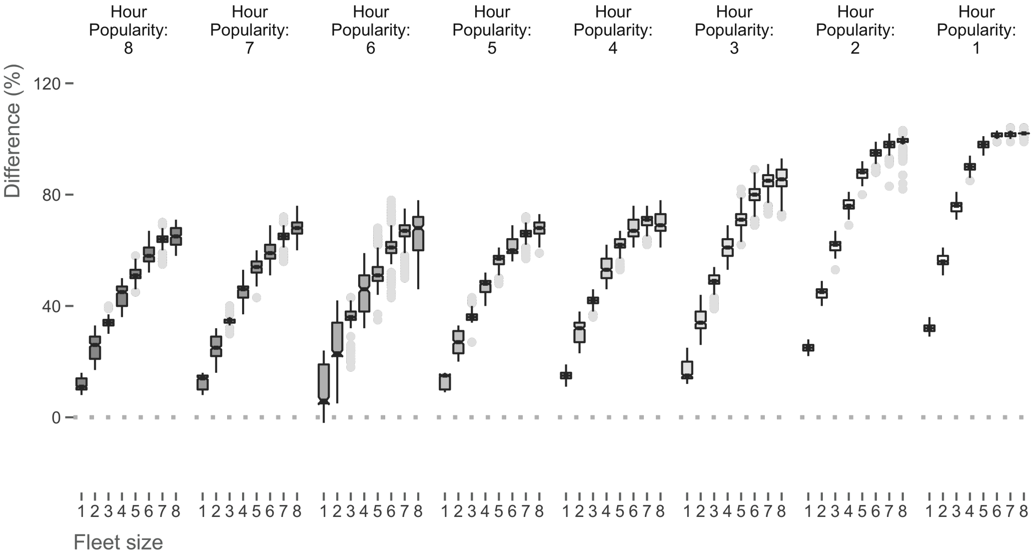

Figure 3.5.10. Percentage-wise surplus coverage by autonomous ferries per time slot compared to manually operated ferries.

While Figure 3.5.9 indicates that autonomous ferries can contribute the most value if introduced at popular hours, Figure 3.5.10 suggests the opposite. Specifically, popular hours include more passengers demanding transportation, which is a natural driver for why an outperforming transportation method—in this case, autonomous ferries—will transport more passengers. If one transforms the outperformance to a percentage of total demanded transportation for the time slot in question, Figure 3.5.10 shows that autonomous ferries provide the greatest coverage for the least popular hours of the week. This outperformance is caused by the fact that autonomous ferries could cover all demand with a sufficient number of ferries. In contrast, manually operated ferries prioritized the more popular hours of the week. Considering that it may not be possible to operate autonomous ferries all hours of the week in reality and that there are other, manually operated transportation alternatives available —such as buses—which also tend to prioritize popular hours of the week, this finding indicates that autonomous ferries may provide great value when used in combination with other transportation alternatives if used during low-demand hours of the day.

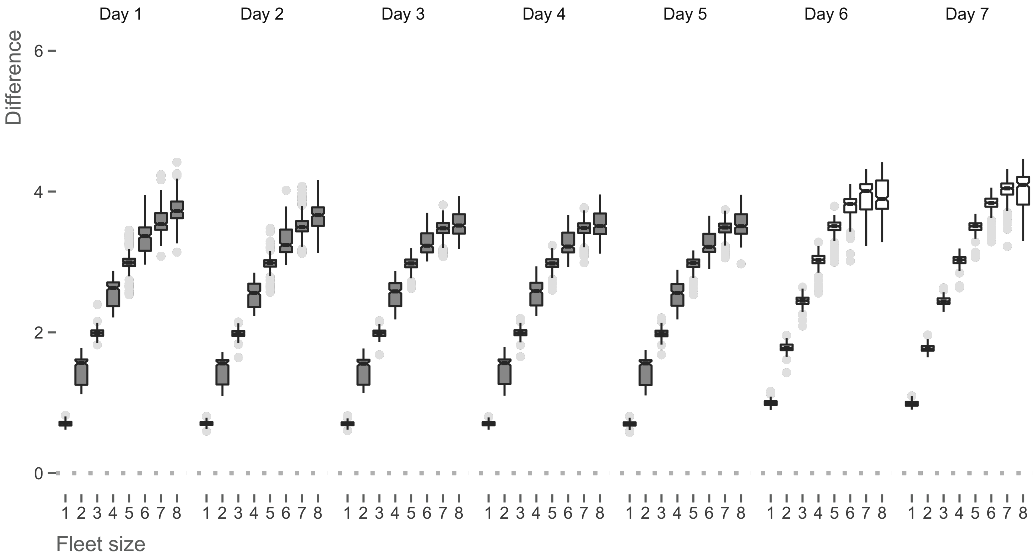

Figure 3.5.11. Surplus coverage by autonomous vehicles across all ferry fleet sizes.

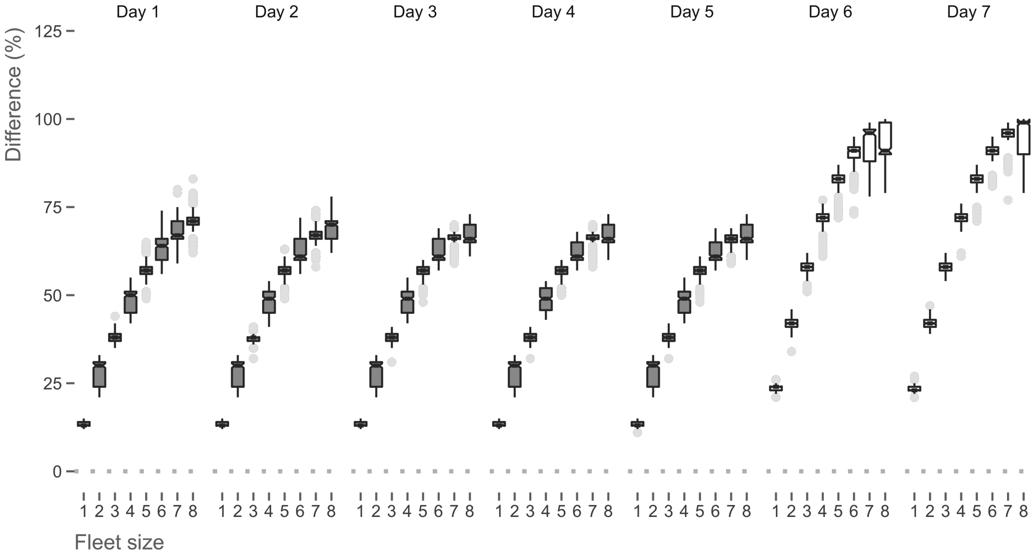

Figure 3.5.12. Surplus coverage by autonomous vehicles, as calculated on an adjusted, *percentage-wise* time slot-basis across all ferry fleet sizes.

Figures 3.5.11 and 3.5.12 strengthen the hypothesis of providing autonomous ferries at times of the week with less demand. Specifically, each day of the week was assigned an ordinal score, with Monday–Friday ranked equally and Saturday–Sunday ranked lower. Just as with hours of the day, autonomous ferries outperform manually operated ferries the most in the percentage of the time slot's total demand when there is lower demand—which is during the weekend from a day-wise perspective. However, autonomous ferries also outperform manually operated ferries the most in terms of the number of excess passengers they provide *despite* the fact that the weekend has less demand for passengers. This outperformance contrasts the hourly perspective, in which Figure 3.5.9 showed a lower outperformance in terms of excess passengers transported for the least demanded hours of the day. Consequently, if few autonomous ferries are available with limited operating time per week, employing autonomous ferries during the weekends may be beneficial when fewer manually operated alternatives are available.

3.6 Concluding Remarks

This article assessed the flexibility of autonomous ferries in transport operations subject to labor regulations, motivated by the case of Bergen, a coastal city in Norway where water buses are currently under consideration. To assess this problem, we formulated two integer linear programming models whose objective function minimizes the total penalty due to the arrival delay from transported passenger groups, the penalty of passenger groups that use alternative transportation, and the number of passengers within each group.

Our computational study showed that the majority of the manually operated ferries are allocated to work during peak hours. Since crews can only work for eight hours before needing a daily rest, these simultaneous work allocations cause a low ferry utilization as the ferries are only operated for some of the day. Consequently, autonomous ferries transported more passenger groups throughout the day than their manually operated counterparts, and this effect happened for all fleet sizes. Moreover, employing only a handful of autonomous ferries that operate 24/7 to cover most of the transportation demand was sufficient.

While fleet utilization could be improved by hiring more crews to operate on time of the week with less demand, it may make little economic sense to do so. Hence, this outperformance by autonomous ferries will likely remain substantial in a practical setting, too. For the test instances presented in this article, autonomous ferries outperformed manually operated ferries throughout the week. However, the severity of it depended on the hour and day of the week. Specifically, autonomous ferries transport more passengers per time slot than manually operated ferries for higher day and hour popularities and bigger fleet sizes.

However, when measuring the outperformance in the percentage of total demand for each time slot, autonomous ferries outperformed the manually operated ferries the most at low-demand hours—at hours and days of the week with the lowest popularity scores. Due to potential practical limitations in the number of operating hours for autonomous ferries and due to other transportation alternatives that will operate the most during high-demand hours, this finding indicates that autonomous ferries may provide significant value when used in combination with other transportation alternatives if used during low-demand hours of the day. Thus, we recommend autonomous ferries to be used from late evenings to early mornings in addition to weekends.

In conclusion, using autonomous ferries in public transportation can significantly improve the passengers' utility, particularly in Norway, where there is a high coast-to-land ratio, road infrastructure is expensive, and a positive outlook on AVs by authorities makes this technology particularly relevant. This finding indicates that there may be a substantial potential for increased revenues by employing autonomous vessels as the number of passengers transported increased in our studies.

Acknowledging the limitation that our article has analyzed a deterministic setting, future work could introduce a stochastic approach to deal with the inherent uncertainty associated to passenger demand and traveling times. Developing new algorithmic approaches to compare against the results of our heuristic and to address larger instances also remains as an interesting avenue for future research. Finally, since our analysis focused in Bergen and used Norwegian labor regulations, performing similar studies in other countries subject to different regulations may add new insights about the introduction of autonomous ferries in passenger transportation.

Acknowledgments

We are thankful to the team of the AUTOLOG project for exciting discussions about the problem studied in this article.

3.A Supporting Functions to the Heuristic

Function 1: Attempt to allocate destination only

```
Definitions:
    ed  earliest time slot the passenger group can depart from the origin.
    ld  latest time slot the passenger group can depart from the origin.
    ea  earliest time slot the passenger group can arrive at the destination.
    la  latest time slot the passenger group can arrive at the destination.
Input: passenger group data, local schedule, local list of allocated passenger groups
Output: local schedule, local list of allocated passenger groups
find first origin in $\{ed, .., ld\}$;
find all open time slots after first origin in $\{ea, ..la\}$;
allocation_successful ← False;
for $\tau$ in open time slots do
    if not allocation_successful then
        find previous allocation's time and location;
        if any previous allocations then
            calculate earliest arrival time to destination if traveling from last allocation;
            if earliest arrival time $\leq \tau$ then
                find the location and time of the next allocation after $\tau$;
                if any next allocations then
                    calculate earliest arrival time to next location if traveling from destination;
                    if earliest arrival to next location is before the scheduled departure from next
                      location then
                        allocate destination on $\tau$ to local_schedule;
                        update the local list of allocated passenger groups;
                        allocation_successful ← True;
                    end
                else
                    allocate destination on $\tau$ to local_schedule;
                    update the local list of allocated passenger groups;
                    allocation_successful ← True;
                end
            end
        end
    end
end
```

Function 1 aims to allocate the destination for a passenger group efficiently, given that its origin has already been assigned within the departure window for the respective passenger group. The function identifies the first available time to depart `origin` within the departure window, $ed, .., ld$. Subsequently, it determines all open time slots after the first origin within the arrival window, $ea, .., la$. After initializing `allocation_successful` as False, the function iterates through the unassigned time slots. For each time slot τ, and until an allocation is made, it evaluates if it is feasible to allocate the destination location at time τ without modifying the already allocated schedule.

Function 2: Attempt to allocate both origin and destination (part 1)

```
Definitions:
    S     set of time slots for the working shift.
    ed    earliest time slot the passenger group can depart from the origin.
    ld    latest time slot the passenger group can depart from the origin.
    ea    earliest time slot the passenger group can arrive at the destination.
    la    latest time slot the passenger group can arrive at the destination.
Input: passenger group data, local schedule, local list of allocated passenger groups
Output: local schedule, local list of allocated passenger groups

find first origin in $\{ed, .., ld\}$;
find all open slots after first origin in $\{ea, .., la\}$;
allocation_successful ← False;

if not allocation_successful then
    departure_interval ← $\{ed, .., ld\} \cap S$;
    arrival_interval ← $\{ea, .., la\} \cap S$;

    if origin is already allocated within departure_interval and destination is already allocated within
      arrival_interval then
        calculate shortest travel time from origin to destination;
        if latest arrival time − earliest departure time ≥ shortest origin–destination travel time then
            allocate passenger group, scheduled at already allocated times for origin and destination;
            update the local list of allocated passenger groups;
            allocation_successful ← True;
        end
    end
end

generate potential departure times, which is the intersect of open time slots and departure_interval;
attempt to allocate both origin and destination—part 2 (Function 3);
```

Function 3: Attempt to allocate both origin and destination (part 2)

Input: `allocation_successful`, passenger group data, local schedule, local list of allocated passenger groups

Output: local schedule, local list of allocated passenger groups

```
for $\tau^D$ in potential departure times do
    if not allocation_ successful then
        find previous allocation's time and location before or on $\tau^D$;
        if any previous allocations then
            calculate earliest arrival time to origin if traveling from the last allocation;
            evaluate if origin is reachable w.r.t. last allocation before origin;
        else
            origin is reachable w.r.t. last allocation before origin;
        end

        find the next allocation's time and location after or on $\tau^D$;
        if any next allocations then
            calculate earliest arrival time to the next location if traveling from the origin;
            evaluate if origin is reachable w.r.t. first allocation after origin;
        else
            origin is reachable w.r.t. first allocation after origin;
        end

        if origin is reachable w.r.t. both previous and next allocations then
            calculate shortest travel time from origin to destination;
            calculate assignable arrival time slots;
            attempt to allocate both origin and destination—part 3 (Function 4);
        end
    end
end
```

Function 4: Attempt to allocate both origin and destination (part 3)

```
Definitions:
  $\tau^D$    potential departure time.
Input:  allocation_successful, passenger group data, local schedule, local list of allocated passenger
        groups
Output:  local schedule, local list of allocated passenger groups
for $\tau^A$ in assignable arrival time slots do
    if not allocation_successful then
        identify time and location of last allocation in $\{\tau^D, .., \tau^A\}$;
    end

    if last allocation exists then
        evaluate if destination is reachable w.r.t. last allocation before destination;
    else
        destination is reachable w.r.t. last allocation before destination;
    end

    if destination is reachable w.r.t. previous allocation then
        if $\tau^A$ is not already allocated to destination then
            if allocation exists after $\tau^A$ then
                identify the time and location of the next allocation after $\tau^A$;
                calculate earliest arrival time from destination to next allocation;
                if earliest arrival time is before or at departure time from next allocation then
                    allocate origin to schedule at $\tau^D$ to local_schedule;
                    allocate destination to schedule at $\tau^A$ to local_schedule;
                    update the local list of allocated passenger groups;
                    allocation_successful ← True;
                end
            else
                allocate origin to schedule at $\tau^D$ to local_schedule;
                allocate destination to schedule at $\tau^A$ to local_schedule;
                update the local list of allocated passenger groups;
                allocation_successful ← True;
            end

        else
            allocate origin to schedule at $\tau^D$ to local_schedule;
            update the local list of allocated passenger groups;
            allocation_successful ← True;
        end
    end
end
```

Functions 2–4 collectively attempt to allocate the passenger group's origin and destination locations within its departure and arrival windows, respectively. In part 1 of the algorithm, the first available `origin` within the passenger group's departure window is identified, and open time slots are found after the first origin in the arrival window. Additionally, `allocation_successful` is initialized as False. The algorithm then checks if the origin is allocated within the departure interval and the destination is allocated within the arrival interval. If so, it calculates the shortest travel time from the origin to the destination location and checks if the time difference between these pre-made allocations is greater than or equal to the shortest travel time. If this condition is met, the origin and destination locations have previously been allocated at times that fit with the passenger group's departure and arrival windows. Hence, the flag variable `allocation_successful` is set to True, and no further allocation attempts occur.

If `allocation_successful` is still False, the allocation attempt continues in part 2. In this part, the algorithm iterates through potential departure times, evaluating the reachability of the origin concerning both the previous and next allocations of time slot τ^D. If the origin is reachable at τ^D, the algorithm calculates the feasible arrival time slots and proceeds to part 3.

In part 3, the algorithm iterates through assignable arrival time slots, evaluating the reachability of the destination concerning the last allocation prior to time slot τ^A. If reachable, the algorithm checks if any allocations exist after the current arrival time slot. If so, the algorithm calculates the earliest arrival time to the following allocation from the passenger group's destination location and checks if this time is before or equal to the departure time from the next allocation. If this condition is met, or if there are no allocations after τ^A, then the passenger group's origin and destination locations are allocated to their respective time slots, and the flag variable is set to True.

Adjusting for Cell Suppression in Commuting Trip Data

Christian Braathen, Inge Thorsen, and Jan Ubøe

Abstract

Maximum entropy methods are used to infer the true trip-distribution matrix in cases where parts of the data are suppressed due to privacy concerns. Large proportions of the suppressed data are found to be inferred correctly when the marginal totals in the trip-distribution are known. Entropy-based approaches are further found to outperform a strategy of ignoring suppressed information in cases with suppressed marginal totals and/or a higher cut-off value of suppressing cell information. Our methods reduce the systematic bias in estimates of the distance deterrence parameter, preventing potentially serious bias in estimates and predictions resulting from standard spatial interaction models. Another useful contribution is to identify what situations an entropy-maximization approach benefits from incorporating information on times series and/or information on distances in the transportation network.

Keywords: trip distribution; suppressed data; entropy; nonlinear programming; mixed integer linear programming

JEL Classification: **R41** Transportation Demand; **R23** Regional Labor Markets; **C15** Statistical Simulation Methods: General; **C61** Optimization techniques; **C44** Operations Research; **C81** Methodology for Collecting, Estimating, and Organizing Microeconomic Data; **C82** Methodology for Collecting, Estimating, and Organizing Macroeconomic Data

4.1 Introduction

Privacy concerns explain why statistical agencies often introduce limitations in releasing data, of which cell suppression is the most commonly used approach for tabular data. This applies, for example, to tabular data on commuting trips between Norwegian census tracts, which are generally suppressed for all combinations of census tracts with less than three commuters. This limitation in data is not introduced for more aggregated subdivisions of the geography, such as the municipality level. Particularly for sparsely populated areas, the cell suppression may involve a considerable number of commuters and potentially lead to seriously biased estimates, for example, on how variations in distance affect the travel demand. Data on commuting trips may not be sensitive on their own. However, a few observations in a cell can be linked to information in other databases to identify individuals and more sensitive attributes.

Based on synthetic data on commuting trips, we discuss three issues related to statistical disclosure limitations. First, we discuss methods to reconstruct the true matrix in cases where we utilize time-series information on commuting trips and distances between specific origins and potential destinations. Second, is the statistical limitation in releasing data, like cell suppression, ignorable for analytical purposes, or does it influence the results substantially regarding biased parameter estimates and predictions? Third, does the common practice of cell suppression represent a reasonable balance between privacy and accuracy? To answer these questions, we apply maximum entropy to time series data. This leads to integer-constrained non-linear optimization problems with many variables. The proposed method is very simple yet surprisingly efficient in reducing parameter biases caused by cell suppression.

Methods to limit disclosure of confidential data represent a challenge for empirical research. As stated in Abowd and Schmutte (2019), “The threats to privacy inherent in the big data era have affected the policies governing statistical agencies” (p.173) and economists “have not fully understood how pernicious suppression bias is” (Abowd and Schmutte, 2016, p.257). Carpenter et al. (2022) demonstrate how imperfect replication of suppressed cell information represents a source of measurement error that may increase standard errors and/or lead to biased parameter estimates. Abowd and Schmutte (2016) recommend that more effort be made to develop other privacy-preserving methods and that a standard should be developed for citing data and for appropriate documentation and discussion of the relevant statistical disclosure limitation (SDL).

The primary motivation for this article is threefold. First, we discuss how to correct adequately for ordinary cell suppression in in commuting trip distribution problems. Second, we discuss the need for the statistical agency to tighten up the usual practice of cell suppression to preserve a sufficient level of confidentiality. Third, we discuss how statistical limitations in releasing data affect the performance of a standard trip distribution model. Are the parameter estimates sensitive to cell suppression procedures, or is suppression ignorable in studying commuting?

The analysis in this article is conducted in terms of spatial interaction, represented by the journey-to-work. Still, we think that the central part of the analysis can be generalized and transferred to other problems. The results are potentially helpful for researchers reflecting on how statistical limitations in releasing data affect their estimation and predictions, and suggesting methods to deal with, for instance, cell suppression. The results are also potentially useful for statistical agencies in designing rules on how to suppress information for the public in preventing sensitive data from being disclosed.

Our results are based on data generated from an agent-based approach. A population of interacting, utility-maximizing agents make labor and housing market decisions in a geography of 12 towns. From the initiation, the inhabitants are randomly assigned characteristics according to frequencies observed in Norwegian data. Hence, the synthetic population in our fictive geography mirrors the Norwegian population in terms of, for instance, age, getting married, the number of children, divorce rates, being retired, death rates, etc. An agent-based approach to providing data for the analysis has the advantage that different aspects concerning confidentiality and trip distributions can be systematically monitored and examined, for example, to avoid issues related to endogeneity and causality.

Section 4.2 provides a review of literature discussing different issues concerning statistical disclosure limitation (SDL). Section 4.3 introduces the geography and explains how the synthetic data are generated from an agent-based approach and, as such, are not exposed to privacy concerns. Section 4.4 demonstrates how an entropy-maximizing approach successfully discloses suppressed information on commuting trips. As reported in Section 4.5, incorporating information on time series in an entropy-maximizing approach significantly discloses suppressed information. In contrast, distance information does not turn out to be relevant in the case with a cut-off suppression value of 3. In Section 4.6, we demonstrate that leaving out information on marginal totals does not lead to a substantial limitation on the possibility of disclosing suppressed information about commuting trips. The experiments further suggest that entropy maximization has the potential of removing a substantial source of bias, also in cases with a higher cut-off value of cell suppression. As reported in Section 4.7, we also find that distances contribute significantly to better fitting in cases with sparse information on the pattern of suppressed information, while time series information no longer contributes. Section 4.8 demonstrates that ignoring the commuters represented by the suppressed cells causes biased estimates of the distance deterrence parameter in a standard doubly constrained gravity model. This bias is, to a large degree, eliminated when the suppressed information is disclosed by entropy-maximization. Finally, concluding remarks are provided in Section 4.9.

4.2 Statistical Disclosure Limitation—a Review of Relevant Literature

The increasing access to microdata has reinforced issues related to privacy and confidentiality. As pointed out by Matthews and Harel (2011), data releasing agencies must account for the fact that privacy is viewed as a fundamental human right by the United Nations. One issue is that some agents may choose to make malicious use of data, and another is that respondents may resist providing honest answers if they can be disclosed by outsiders (Matthews and Harel, 2011).

On the other hand, "more powerful computers and advances in algorithms such as machine learning have led to an explosion in the usefulness of data" (Jones and Tonetti, 2020, p.2819). Hence, confidentiality concerns must be balanced against the utility of reaching new insight with potentially important policy implications. In an optimization context, the released data should minimize the loss of information necessary to preserve the privacy of the individuals in the database. Abowd and Schmutte (2019) refer to literature proving that there is no free lunch, which means that any publishing of useful statistical summaries gives a loss of privacy. At the same time, suppressing information due to privacy concerns introduces a measurement error to the data. This may, in particular, influence the analysis in cases where different categories of workers and industries are introduced for a spatially disaggregated specification of a geography in a sparsely populated, rural area. Carpenter et al. (2022) point out that there is a trade-off between the bias that can be expected from using an aggregate representation of a variable and the bias resulting from more suppressed information if a more disaggregated approach is followed.

4.2.1 Defining and Measuring Privacy and Accuracy

Consider first different concepts concerning privacy. Willenborg and De Waal (2012) distinguish between the risk of re-identification, which relates to the identification of an individual, and a predictive disclosure, which is about identifying values of a sensitive attribute for an individual. Abowd and Schmutte (2016) denote the first as an identity disclosure, while G. Duncan and Lambert (1989) distinguish between attribute disclosure and inferential disclosure. The former relates to obtaining reliable information due to linking attacks. At the same time, the latter occurs when reliable information is reached even without linking to observations from another database. G. Duncan and Lambert (1989) discuss population, or model, disclosure, which is about reaching confidential information about a population using a model based on the released microdata. In more recent literature—Abowd and Schmutte (2016)—this kind of disclosure is treated probabilistically, reflecting the ability that the published data can be used to identify individuals or attributes with a substantially higher probability. As pointed out by Abowd and Schmutte (2016), this applies—for instance—in approaches that use the randomized response to sensitive questions.

In addressing the measuring and definitions of privacy, Shlomo (2018), Chetty and Friedman (2019), and Abowd and Schmutte (2019) focus on the concept of differential privacy, building on ideas from for example Dwork et al. (2006). This concept is based on the introduction of a parameter ϵ, which measures "the maximum difference in the log odds of observing any statistic across similar databases" (Abowd and Schmutte, 2019, p.175) and is distinguished by whether confidential data are included or not. Hence, the privacy loss is limited by defining ϵ as an upper bound of the likelihood ratio (Chetty and Friedman, 2019), corresponding to a maximum risk accepted for a released statistic for preserving privacy. This parameter reflects the possibility that potential intruders can identify sensitive individual information from published statistics. In other words, the differential privacy is related to the loglikelihood ratio defined by the likelihood that the released parameter estimate stems from a specific dataset rather than a dataset that differs by one observation. If this likelihood ratio falls, for example, due to noise infusion, it will be more difficult to distinguish between two marginally different datasets. Hence, this will increase the likelihood that privacy is preserved even if parameter estimates based on a small sample of individuals are released.

Many different methods to measure privacy was introduced in the 80s and the 90s—see, for instance, Matthews and Harel (2011) for a review. Sweeney (2002) introduced the measure

k-anonymity, achieved if a specific combination of individual characteristics appears at least k times in a table. Grouping and/or cell suppression are techniques that can be introduced to achieve k-anonymity. A potential problem with this measure is that disclosures may find a place if there is a lack of diversity in the occurrence of sensitive attributes and if a potential intruder has some background information on specific individuals and/or attributes. The literature provides suggestions on overcoming such problems—see, for instance, Matthews and Harel (2011). They, in addition, discuss the presence of inferential disclosure, where information on some attributes, like gender, age, occupation, and region (explicit identifiers), can be combined with a more sensitive variable to build a regression model, reaching a predictive distribution of the values of the sensitive variable. As claimed by Abowd and Schmutte (2016), it is, from a probabilistic perspective, in general, impossible to release data without compromising confidentiality.

Another issue related to statistical inference is cases with sporadic observations. Rajasekaran et al. (2009) put forward the idea that observations far from the mean have a high risk of being disclosed and thereby introduce a rationale for suppressing cells with few observations. This can, for example, be the case for commuting between zones involving a long distance.

4.2.2 Balancing Confidentiality and the Utility of Research

As stated above, a trade-off between privacy protection and accuracy represents an optimization problem. Abowd and Schmutte (2019) argue that the accuracy of published statistics can be seen as non-rival and non-excluding in consumption and hence considered a public good. Similarly, it is argued that the differential privacy parameter ϵ reflects a social issue and that an SDL procedure gives the entire population the same protection against privacy loss. This means that privacy protection is also strictly non-rival and can be considered a public good—see the discussion provided by Jones and Tonetti (2020). Since statistical accuracy and privacy are public goods, their optimal levels are a social choice (Abowd and Schmutte, 2019). Abowd and Schmutte (2019) provide a normative approach, and they claim that making their framework practical calls for more sophisticated models of production possibilities and better models and measures concerning the demand for privacy and accuracy (Abowd and Schmutte, 2019). They also claim that statistical agencies should aim at solutions where the marginal costs are equal to the marginal benefits of statistical disclosure limitation procedures. However, “statistical agencies are not yet using formal privacy protection systems”, (Abowd and Schmutte, 2019, p.203), and not much research is provided on such issues.

Abowd and Schmutte (2016) point out that the effects of SDL on the results are more severe when the analysis aims at explaining the situation for a specific sub-population that is in particular exposed to the confidentiality issue. They further define SDL to be ignorable if "the analysis can recover the estimates of interest and make correct inferences using the published data without explicitly accounting for SDL" (Abowd and Schmutte, 2016, p.223). In general, SDL may cause a severe bias in a rural-urban dimension, reflecting the likely event (Carpenter et al., 2022) that less-populous geographies have a higher share of suppressed cells. Another perspective concerning the trade-off between privacy and the usefulness of data is related to the property rights for data. Jones and Tonetti (2020) address the possibility that consumers' privacy is not adequately accounted for if firms own the data. At the same time, the nonrival character means potentially substantial social gains in making data broadly accessible.

4.2.3 Alternative Ways of Limiting Disclosure

One obvious step to avoid disclosure of private information is removing data that directly allows identification, like name and home address. However, this is generally insufficient to maintain the individual's privacy, as demonstrated in Sweeney (2002). Hackers may, for instance, combine different databases to reach sensitive information, and in general, a small number of demographic attributes allows the identification of an individual (Abowd and Schmutte, 2016). A database of location information, such as an origin-destination matrix of commuting, can serve as an example of data that can be used to disclose more sensitive individual information from other databases.

Matthews and Harel (2011) provide a review of methods employed for maintaining the privacy of publicly released data, and a comprehensive review can be found in G. T. Duncan et al. (2011). Entering into details on different methods is beyond the scope of this article. However, the following list of methods, based on Matthews and Harel (2011), may serve as a valuable backdrop for the discussion to follow:

1. Basic methods:

   - *limitation of detail*, in which variables can be recorded into intervals, and/or categories can be collapsed together;
   - *top/bottom coding*, in which a variable's largest or smallest value is limited;

- *suppression*, in which cells with too few observations in a contingency table are not released to the public but replaced with missing values;
- *rounding*, in which each observation is rounded up or down to the nearest multiple of the rounding base, and where rounding up or down is decided upon randomly; and
- *addition of noise*, in which noise can be added to the data to prevent identification through linkages.

2. Sampling: If only a microdata sample is released, potential intruders cannot be sure that a unique match from another dataset identifies an individual.

3. Matrix masking: Appropriate conformable matrices are released. This includes noise addition, sampling, suppressing sensitive variables or cells, adding simulated data, and all exceptional cases of matrix masking. Another option is to use randomized responses to gain sensitive information, by introducing with a specific probability a trivial question as an alternative to the sensitive question. The sensitive information then remains private even in cases of identification of individuals.

4. Data swapping: According to Abowd and Schmutte (2016), "Data swapping is the practice of switching the values of a selected set of attributes for one data record with the values reported in another record" (p.230). The data-swapping only needs to be done on sensitive variables by changing units between cells.

5. Synthetic data: The idea of generating synthetic data for statistical disclosure limitations was introduced by Rubin (1993). Abowd and Schmutte (2016) introduce the possibility that synthetic data are validated by the data providers on the actual confidential data. As explained in Section 4.3, we generate synthetic data to discuss challenges and corrections for suppressed data in a commuting context.

6. Aggregation and other selected methods: Slicing complete data into groups involving a smaller number of variables, while microaggregation is based on creating new records by averaging at least three original records. Geographic units can be aggregated, data on occupation and industry can be released in broad categories, and incomes can be reported in bins (Abowd and Schmutte, 2016).

7. Micro-agglomeration, substitution, subsampling, and calibration: A combination of different statistical disclosure techniques, proceeding in steps and ending by calibrating the

released data such that specific estimates based on released data match the estimates from the observed data.

This list of methods to deal with SDL refers to the use of microdata. If such methods are applied to microdata, they also represent a source of error in tabular data. This is particularly unfortunate if the analysis focuses on issues sensitive to observations in cells representing choices made by a few individuals in the population. According to Abowd and Schmutte (2016), the most common method to deal with sensitive cells in a table is suppression, while randomized rounding is also used in many cases. Concerning suppression, Salazar-Gonzalez (2004) proposes to suppress the cells that minimize the loss of information while guaranteeing suppression level requirements. He suggests four different mathematical models for selecting which cells to suppress. Abowd and Schmutte (2016) further claim that adding noise to the microdata may be a preferred alternative to suppression. Noise infusion means that variance will be added to published data while reducing bias. Chetty and Friedman (2019) discuss how the risk of privacy loss can be reduced by adding noise to parameter estimates.

Abowd and Schmutte (2016) also explain the need for complementary suppression for tabular data. This is, for instance, due to the possibility that users can deduce values of sensitive cells from information on, for example, marginal sums in the table. Statistics Norway demands that at least three units have to underlie a total value to be published in a cell to avoid the risk of a residual disclosure. In many countries, this is not considered sufficient, restricting that the three main contributors to the cell total represent at least 80% of the total sum. This is a case of the so-called (n, k) dominance rule, suppressing the cell if the n largest units contribute more than k% of the total sum. This dominance rule is primarily introduced to keep vital information hidden for competing firms, for instance, concerning variables like production, sales, wage costs, import, export, etc. Still, this information may be available from other sources, and official statistics will probably not be the first source of information where competing firms search for relevant information.

A challenge researchers face is that data publishers generally do not disclose their complementary suppression methods. However, Abowd and Schmutte (2016) claim that data-providing agencies are becoming willing to use noise-infused methods in producing data tables. Adding noise to statistics generated from a database belongs to a class of approaches called matrix mechanisms, which are elaborated in Li et al. (2015). According to Carpenter et al. (2022), US statistical agencies use cell suppression and noise infusion, "with some published cell values perturbed by a random noise multiplier, to prevent the disclosure of individual business establishment information" (Carpenter et al., 2022, p.58).

An approach with noise infusion may also be performed by adding noise to the estimates from the regression based on a small sample. This technique is based on the differential privacy concept discussed in Subsection 4.2.1. Chetty and Friedman (2019) compare such a procedure to a standard cell suppression approach. Their comparison can be summarized in terms of the following three dimensions:

Privacy loss Chetty and Friedman (2019) claim that for most noise infusion approaches, the risk of privacy loss is substantially reduced compared to cell suppression.

Statistical bias The parameters underlying the random noise infusion are publicly known (Chetty & Friedman, 2019). Hence, unbiased parameters can be reached, as opposed to what, in general, is following from the measurement errors resulting from count-based suppression approaches.

Statistical precision Estimates based on noise infusion are less precise than those following cell suppression methods. According to Chetty and Friedman (2019), this is the key drawback of noise infusion and the primary concern of most researchers.

4.2.4 Accounting for Statistical Disclosure Limitation Methods

In many cases, disclosure limitations will not be expected to affect the results of empirical analysis. According to Abowd and Schmutte (2016), modifications due to confidentiality concerns are generally modest relative to other more serious data quality problems, like reporting errors and missing items.

Suppose methods for statistical disclosure limitations prove to be non-ignorable in terms of affecting the results. In that case, they should be explicitly adjusted for in the analysis to avoid biased estimates of parameter values and/or the corresponding variances. Such adjustments call for knowledge of the SDL methods used by the data publisher or methods that can recover the SDL parameters from prior information and the released data. From a researcher's point of view, it would be valuable if details on the appropriate SDL method were made public by the data-releasing agencies.

Abowd and Schmutte (2016) provide two examples of such methods. One is an approach with randomized responses to sensitive issues, while the other is top coding of incomes, censoring incomes above a specific threshold. In both cases, the SDL method obviously should be accounted for in analyzing the data—see Abowd and Schmutte (2016) for a discussion. Abowd and Schmutte (2016) also discuss how SDL methods may lead to attenuated parameter estimates and underestimated standard errors in linear regression models. They refer to analyses addressing the closely related biases resulting from other sources of missing data, like missing responses to specific questions in a survey. Abowd and Schmutte (2016) provide a discussion of how the bias resulting from SDL can be corrected. One approach is to use the information on the noise variance in a case where this is the relevant SDL and then correct the bias analytically. As another approach, it is known that finding an appropriate instrument may be an appealing solution to deal with measurement errors (Abowd and Schmutte, 2016; Carpenter et al., 2022).

Abowd and Schmutte (2019) distinguish between a reconstruction attack and a re-identification attack. The former is about building a copy of a confidential database based on statistics produced and published from this database. If linearly dependent statistics are available, there may be a substantial potential for reconstructing hidden variables, which is a data breach. Marginal totals from a contingency table are, i.e., linear statistics that can be used for reconstruction. Re-identification involves linkage to information from other external databases, deterministically or probabilistically, see for instance Airoldi et al. (2011).

Carpenter et al. (2022) discuss an increasing demand for private data sources. Private agents may provide data where suppressed cells are estimated, for example, through the availability of new data, like online business directories, underpinned by increased computational power. According to Carpenter et al. (2022) and Abowd and Schmutte (2019), this may lead to privacy concerns and to a higher percentage of suppressed cells. Hence, an increasing interest can be expected in how cell suppression can invalidate statistical inference and how this should be treated analytically. Carpenter et al. (2022) claim that most journal articles in economic literature do not document the methods used in estimation based on suppressed cell data sets. One possible technique, mentioned by Carpenter et al. (2022), is an iterative proportional fitting procedure, which is "an algorithm for estimating cell values of a contingency table such that the totals remain fixed and the estimated table decomposes into an outer product" (Carpenter et al., 2022, p.61). This technique has been used (Bartik et al., 2018), resulting in estimates labeled WholeData, which are free to researchers. Carpenter et al. (2022) use WholeData to examine measurement errors resulting from suppressed cell data sets, and they claim that many firms offering purchasable data use nonpublic algorithms to provide suppressed cell data estimates. Guldmann (2013) provides a review of methods to deal with suppression issues. According to

Carpenter et al. (2022), "methods doubtless vary in their success, but none address the larger question of how much error remains and how that error influences economic and statistical significance" (p.62).

4.3 Generating a Synthetic Population and Data for the Analysis

As noted in Section 4.2.3, one possible statistical disclosure limitation method is generating a synthetic population. Commensurate with the fundamental problem being discussed in this article, we do the analysis based on a synthetic population. Hence, apart from reaching insights on spatial interaction behavior that may contribute to identifying population subgroups, we are not faced with confidentiality issues.

A synthetic population can be generated through agent-based modeling. As pointed out by A. Wilson (2010), this is represented by a system where the individual agents make decisions according to probabilistic rules of behavior. The observed spatial pattern results from decentralized decision-making, as demonstrated in for example Page (1999), Anas (1983a) and Irwin (2010). The construction of such a pattern is essentially reflecting two separate steps. The first specifies the agents' geography and population, while the second introduces the labor and housing market conditions that define the opportunity set of the utility-maximizing agents. For an example of how such a modeling framework is parameterized and applied, see Gholami et al. (2022).

4.3.1 Specifying the Geography and Generating a Population

The geography is represented by the 12-node system illustrated in Figure 4.3.1. There are three clusters of towns. The towns A, B, C, and D are located on another side of a topographical barrier than the other towns. There are short distances between the towns within each cluster but relatively long distances across the different clusters. This means that a relatively high number of cells in the matrix can be expected to involve a low number of commuters, leaving an imminent risk of identification disclosures.

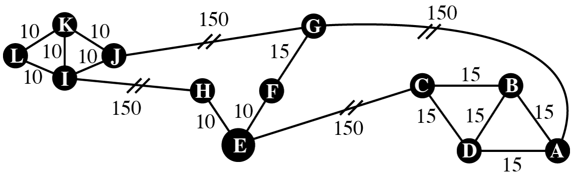

Figure 4.3.1. The geography of towns. The numbers marked in the figure represent the kilometers' distances for the corresponding segments in the road transportation network.

For technical convenience, similar to the construction in McArthur et al. (2010) and McArthur et al. (2012), the generation of a population is initiated with a large number of fifteen-year-old agents, who are distributed between the 12 towns and by gender according to given probabilities. The population members then interact according to rules based on Norwegian statistical data. This involves probabilities of getting married, having children, being divorced, being retired, dying at a specific age, etc. More details of the demographic rules underlying the formation of a population can be found in McArthur et al. (2010) and McArthur et al. (2012), who consider a simple two-node geography.

The interaction succeeding the initialization goes on for 300 years. After this period, there are no more traces of the initial population, and we are left with a population that mimics the Norwegian in a demographic sense. Our microsimulation experiments start in year 300, resulting, for instance, in a pattern of commuting flows corresponding to individual decisions determined by preferences and budget considerations. The preferences for each agent are specified as a Cobb-Douglas function of their housing consumption and the consumption of other goods and services. This involves a parameter reflecting the relative weight put on these aggregates in the preferences. For each person, the value of this parameter α_i is drawn randomly from a normal distribution with $\alpha_i \in (0.3, 0.7)$.

The total consumption and the utility level depend on an individual's lifetime earnings, which mirror the labor and housing market situation in the different towns of the region. Workers apply for jobs that contribute to increasing the utility level of the household. There are three categories of workers and three categories of related jobs. Within each group, the workers are assumed to be homogeneous regarding job qualifications. Applicants are randomly selected for vacant positions, and a job offer means they can stay in this position until they retire.

If a worker accepts a job offer from a firm located in another town than his/her current residence, there are two possible forms of spatial interaction response. One is to move to the town hosting the new employer, and the other is commuting. The chosen alternative maximizes the sum of utilities of the spouses.

In line with economic base modeling, employment in basic sector firms is assumed to be exogenously given, while local sector firms serve the population in the region. There is a cluster of local sector firms in Town E, which is the central business district, attracting customers from other towns in the region (Gjestland et al., 2006). This generates commuting flows between the towns, as does the friction represented by the heterogeneity of job and worker categories.

Due to a hypothesis of agglomeration economies, wages are initially assumed to be higher in a regional center, town E, than in the other towns of the region. For the years to follow, spatial wage disparities may follow from local Phillips style mechanisms—see McArthur et al. (2010) and McArthur et al. (2012).

Spatial disparities in housing prices may strongly influence the decision to move or commute as a response to a shift in the job location. The housing demand is represented by first-price sealed bid auctions, where all bids are submitted simultaneously, and the highest bid wins the auction. Our model has a pool of houses for sale, and agents make a random check into this pool to see if they may join the pool of potential bidders. All agents in the pool scan the pool of houses for sale, searching for the one that provides them with the highest utility. A sale is completed if a seller finds the winning bid exceeds his or her reservation price. If the sale is not completed, the house remains in the pool of houses for sale. For more technical details on this bidding procedure, see Gholami et al. (2022).

The initial housing supply is assumed to be provided by a regional government planning entity. In our model, high housing prices in the previous year serve as an incentive and explanation for increased building frequencies in the current year.

4.3.2 Adding Workers to the System

As stated in Section 4.3.1, the geography is constructed to reach a solution where a relatively high number of cells in the matrix can involve a low number of commuters. In our micro-simulations, this can be reinforced by the specification of wages, housing prices, and other parameters that influence individual spatial interaction decisions. Hence, the case we consider is constructed to reach a state inducing confidentiality issues.

Through the generation of the synthetic population, we know the state of each agent at any time, represented by their family relations, preferences, residential location, and work location. This means that the corresponding commuting flow pattern results from utility-maximizing agents' decisions. To get a higher diversity of low numbers in the cells of the trip distribution matrix, we systematically add some long-distance commuters to the system. This further contributes to introducing exciting information regarding the time-series properties of the commuting pattern. As we will demonstrate in Section 4.5, time-series data proves to help disclose information suppressed by the data publisher.

Our relatively small system, with only 12 towns, has relatively few cells where the value changes considerably over time. Adding a few long-distance commuters extends the potential to study the relevance of time-series effects. We can, for instance, add 1 or 2 workers to all the cells where 0 or 1 workers are observed in the commuting trip matrix following directly from our micro-simulations for year 300. These workers are assumed to keep their initial combination of residence and job location for the entire period under study. They are, of course, not rational agents making utility-maximizing decisions, and they make up an insignificant proportion of the otherwise rational synthetic population. This approach resembles the introduction of so-called noise traders in finance, that is, traders making irrational decisions, trading by random, deviating from market averages. Adding a few irrational, uninformed workers improves the potential of studying the effect of dynamics and transitions in adjusting for statistical disclosure limitations.

4.3.3 Non-Suppressed and Suppressed Trip-Distributions

For each year, there is a total of l_i workers in town i, e_i working places, and we let T_{ij} denote the number of workers who live in town i and work in town j. The actual trip-distribution matrix is a 12×12 matrix. An example of an actual trip-distribution is shown in Table 4.3.1, which refers to year 17 of a 20-year long time period following from the initialization of our agent-based generation of the population. Our approach works particularly well for year 17, which makes this year appropriate in clarifying what we aim to achieve by the procedure described in Section 4.4.2. The sample results from the procedure described in Sections 4.3.1 and 4.3.2. In the following the actual trip-distribution is referred to as the non-suppressed matrix.

The matrix of commuting trips in Table 4.3.1 reflects the spatial configuration shown in Figure 4.3.1, with a long distance between the three clusters of towns. There are many cells with a low number of commuters, but these workers do not make up a high proportion of the total number of commuters. In the suppressed trip distribution, all the numbers that are encircled in Table 4.3.1 are substituted by a zero entry.

A zero entry in a suppressed matrix of commuting trips, e.g., Table 4.3.1, means that the actual value can be 0,1, or 2, assuming a cut-off value of 3. The basic approach taken in the article is that the total number of workers and working places in each town are public, and our challenge is hence to offer a prediction of the actual trip distribution based on the suppressed data and the known marginal totals from the non-suppressed matrix of commuting trips in Table 4.3.1.

Table 4.3.1 provides the results of our disclosure procedure in a case with a cut-off value of 3. To demonstrate how discrepancies cancel each other out to reach a balanced trip matrix, the example in Table 4.3.1 corresponds to a case where our optimization procedure provided a close to perfect reproduction of the true trip distribution. This stems from using the objective function presented in Section 4.5.2, and the results for year 17 in Table 4.5.1.

Table 4.3.1. Trip-distribution for year 17. Numbers encircled by a gray circle were suppressed and correctly predicted using the objective function presented in Section 4.5.2, while numbers encircled by a black circle were suppressed but incorrectly predicted.

	A	B	C	D	E	F	G	H	I	J	K	L
A	534	96	65	98	89	3	2	2	1	1	1	1
B	107	372	112	100	91	1	3	0	1	1	1	1
C	93	92	390	104	114	5	1	4	1	3	1	0
D	108	93	142	339	104	1	2	1	1	1	1	1
E	5	3	3	5	1470	190	144	288	3	1	3	1
F	2	2	4	2	623	228	103	42	2	1	3	2
G	2	2	1	1	433	26	206	20	1	1	2	1
H	2	3	2	4	555	45	61	231	2	1	1	1
I	1	1	1	1	114	1	3	2	294	130	111	121
J	2	1	2	1	109	4	2	2	138	464	145	69
K	1	1	1	1	83	3	1	4	107	102	348	102
L	1	1	1	1	114	2	1	8	141	96	122	293

It follows from Table 4.3.1, that we have a near-perfect match, the only difference is the four entries encircled in black. It is straightforward to see that the discrepancies cancel each other out regarding the marginal sums of rows and columns. The prediction of 1 rather than 0 commuters from town D to town J is balanced by the prediction of 1 rather than two commuters from town D to town L. Hence, the row sum will be left unchanged. Column-wise, the prediction of 1 commuter from town D into town J shows that the predicted commuting from town F to town D is one rather than the true value of 2 commuters.

4.4 Entropy

Several authors have studied entropy methods in connection with suppressed data, but to the best of our knowledge, this article is the first to discuss such methods in the context of trip distribution.

A. Wilson (2010) claims that entropy-maximizing models are appropriate for estimating missing data, for instance, when row- and column sums are known. Airoldi et al. (2011) propose an entropy metric for assessing how the risk of trail disclosure depends on the distribution of how people visit different sets of locations that may share their information for linkage purposes. Through case-based and controlled experiments, Airoldi et al. (2011) demonstrate that the entropy metric effectively estimates the risk of trail disclosures, suggesting that low entropy systems correlate with high re-identifiability. In general, entropy-based measures have been used to estimate information loss (Willenborg and De Waal, 2012) and also for estimating disclosure risk (Bezzi, 2007). Rodrigues (2016) uses the maximum entropy principle to determine stochastic properties of data where uncertainty results from suppression or other kinds of measurement errors. In the specification of the optimization problem, Rodrigues (2016) accounts for the possibility that there are correlations in the probability distribution of missing observations. Such correlations follow accounting identities, like the marginal totals of a trip distribution problem. Majeed and Lee (2020) focus on the users' privacy in social networks. They estimate the level of uncertainty for disclosure by computing the entropy of sensitive attribute values, with high values of entropy meaning that there is a substantial potential for protecting the privacy of the users of social networks. González-Vidal et al. (2020) demonstrate that a framework with Bayesian maximizing entropy is generally very well suited to deal with missing value imputation in Internet of Things applications.

Entropy considerations date back to the classical works of Boltzmann. Boltzmann's idea was to count how many ways a given macro-state can be realized in terms of randomly selected micro-states, and he assumed that the most likely macro-state was the one that could be realized in the largest number of ways. Shannon entropy, also called information uncertainty, works similarly. In both cases, the entropy H is defined by

$$H = -\sum_{i=1}^{N} P(x_i)\log(P(x_i))$$

where $x_1, x_2, ..., x_N$ are the possible microstates and for $i = 1, \dots, N$, $P(x_i)$ is the probabilities that the microstate x_i is selected. In the context of trip distributions, the agents select where to live and where to work. In a system with n towns, we get $N = n^2$ possible microstates. If we let π_{ij} denote the probability that an agent selects to live in town i and work in town i, the entropy is

$$H = -\sum_{i,j \in I} \pi_{ij} \log(\pi_{ij})$$

where $I = \{1, 2, \dots, n\}$ is the index set of the n towns. If M is the total number of agents living in the system, we expect $T_{ij} = M\pi_{ij}$, and without loss of generality, we can define $\pi_{ij} = \frac{T_{ij}}{M}$. If we rewrite the entropy noting that (by definition) $\sum_{i,j \in I} T_{ij} = M$, we get

$$\begin{aligned} H &= -\sum_{i,j \in I} \frac{T_{ij}}{M} \log\left(\frac{T_{ij}}{M}\right) = -\frac{1}{M}\sum_{i,j \in I} T_{ij} \log(T_{ij}) + \frac{1}{M}\sum_{i,j \in I} T_{ij} \log(M) \\ &= -\frac{1}{M}\sum_{i,j \in I} T_{ij} \log(T_{ij}) + \log(M) \end{aligned}$$

We hence see that in a system with a fixed number, M, of agents, maximum entropy is obtained when $-\sum_{i,j \in I} T_{ij} \log(T_{ij})$ is as large as possible.

4.4.1 Solution to Constrained Maximum Entropy Problems

To infer the most likely trip distribution under marginal constraints, we search for the solution to the following optimization problem:

$$\max_{T_{ij}} -\sum_{i,j \in I} T_{ij} \log(T_{ij}) \quad s.t. \quad \sum_{i \in I} T_{ij} = e_j, \quad \sum_{j \in I} T_{ij} = l_i$$

This problem is easy enough to admit a closed-form solution, and it is straightforward to see that maximum entropy is obtained when $T_{ij} = A_i B_j$, where $A_1, \dots, A_n, B_1, \dots, B_n$ are constants. Numerical values for these constants can be found from the balancing constraints

$$\sum_{i\in I} A_i B_j = e_j \qquad \sum_{j\in I} A_i B_j = l_i$$

Efficient numerical algorithms—such as the Bregman algorithm (Bregman, 1967)—can be used to find these constants. See Sen and Smith (1995) for more. The solution to the maximum entropy problem coincides with the expectations in a Chi-square table under independence.

The problem becomes more interesting if we impose further constraints. In the trip-distribution problem, it is natural to assume a generalized cost c_{ij} when an agent commutes between the origin i and the destination j. Such generalized costs can include start-up costs, driving distance, travel time, etc. Assuming that the total generalized commuting cost C is given, we might consider a modified maximum entropy problem, i.e.

$$\max_{T_{ij}} - \sum_{i,j\in I} T_{ij}\log(T_{ij}) \quad s.t. \quad \sum_{i\in I} T_{ij} = e_j, \quad \sum_{j\in I} T_{ij} = l_i, \quad \sum_{i,j\in I} T_{ij}c_{ij} = C$$

This modified problem, too, admits a closed form solution: There exists constants $A_1, \ldots, A_n, B_1, \ldots, B_n$, and a unique constant β such that

$$T_{ij} = A_i B_j e^{-\beta c_{ij}} \tag{4.1}$$

The solution to the modified maximum entropy problem is a multinomial logit model. It is interesting to note that the solution coincides with the solution to the random utility problem, i.e., the resulting distribution we get when agents select origins and destinations subject to random utility maximization, McFadden (1974) and Train (2003).

The model in (4.1), can be derived in several additional ways, it can be derived from maximum entropy considerations—see A. G. Wilson (1967), Anas (1983b), and Erlander and Stewart (1990). It is the solution of the maximum utility problem Erlander and Stewart (1990) and a consequence of probabilistic cost efficiency Erlander (2010). In addition, there exist several other approaches, such as Mattson and Weibull (2002) and Matějka and McKay (2015), leading to the same model. Contrary to the many economics models chosen for analytical convenience, the approach leading to (4.1) is hence firmly anchored in statistical theory.

4.4.2 Inferring Suppressed Data Using Maximum Entropy Methods

In the previous section, we have seen that a maximum entropy approach produces models consistent with models widely used in the literature on trip distribution. To proceed to cases with suppressed data, we let $\mathcal{S}$ denote the index set of the suppressed entries. We then consider a modified maximum entropy problem:

$$\max_{T_{ij}} - \sum_{(i,j)\in\mathcal{S}} T_{ij}\log(T_{ij}) \quad s.t. \quad \sum_{i\in I} T_{ij} = e_j, \quad \sum_{j\in I} T_{ij} = l_i$$

and where T_{ij} is a free optimization variable for all suppressed variables and equals $T_{ij}^{\text{non-suppressed}}$ otherwise.

We fix all the non-suppressed entries and use entropy maximization to infer the suppressed values. In problems of this kind we use the convention that $x\log(x) = 0$ when $x = 0$, which comes as a natural consequence of the continuous limit $\lim_{x\to 0^+} x\log(x) = 0$. Computer software like AMPL can easily handle integer-constrained optimization problems of this sort.

In general, we will report the fit in terms of accuracy, which we define by

$$\text{accuracy} = \frac{\#\text{ correctly predicted values}}{\#\text{ suppressed values}}.$$

In our analysis, we will also report the fit in terms of SRMSE over the suppressed entries, which we define as:

$$SRMSE = \sqrt{\frac{\sum\limits_{(i,j)\in\mathcal{S}} (T_{ij} - T_{ij}^{\text{non-suppressed}})^2}{|\mathcal{S}|}} \Bigg/ \frac{\sum\limits_{(i,j)\in\mathcal{S}} T_{ij}^{\text{non-suppressed}}}{|\mathcal{S}|}$$

i.e., we compare the mean square deviation over the suppressed entries with the mean of the actual values that have been suppressed. Here $\mathcal{S}$ are the suppressed entries, and T_{ij} is the solution to our modified maximum entropy problem. The SRMSE is a dimension-free quantity that allows us to compare cases where the number of suppressed values is very different.

The example considered in Table 4.3.1 was selected to clarify the prediction procedure. To examine the performance in more general cases, based on a synthetic population for the 12-node geography illustrated in Figure 4.3.1, we started with a pure entropy maximization approach without accounting for additional information on the system. This is done for 20 years, and we generated a time series of non-suppressed matrices from year 1 to year 20.

We believe our procedure is neutral in that there is no part of the construction we expect would favor maximum entropy methods. In the time-series approach discussed in the next section, however, we wanted a data set where several cells flip between suppressed and public over time. As the original data set contained few such cases, we traced the cells with 0 or 1 commuters in the first year and added either one or zero non-optimizing workers in each cell. This worker kept the initial combination of residence and job location for the entire period. As explained in Section 4.3.2, we experimented with three other ways of adding workers but settled with the procedure detailed above.

Entropy maximization, in general, works fine in disclosing suppressed information on commuting trips. As an average over the 20-year-long period, the accuracy is found to be 0.7537, which means that around 3 out of 4 suppressed cells are perfectly disclosed by maximizing the entropy. In comparing the number of disclosed commuters in the suppressed cells to the known actual numbers, an average value of 0.3913 was reached for the SRMSE. For the performance of this pure entropy maximization approach for specific years over the period, see the first five columns of Table 4.5.1.

4.5 Extension to Time Series

The results presented in the previous section demonstrated that entropy maximization might perform well in disclosing information suppressed in the matrix of commuting trips. One interesting question is whether information on the suppressed trip distribution at several consecutive points in time can be utilized to give even more accurate estimates of the suppressed information.

Our core model extends trivially to the time series case. We let $\mathcal{T}$ denote the index set of times, and for each $t \in \mathcal{T}$ we let $\mathcal{V}$ denote the index set of the entries in the trip-distribution that are suppressed at time t. We then consider

$$\max_{T_{ijt}} - \sum_{(i,j,t) \in \mathcal{V}} T_{ijt} \log(T_{ijt}) \quad s.t. \quad \sum_{i \in I} T_{ijt} = e_{jt} \forall j \in I, t \in \mathcal{T}, \quad \sum_{j \in I} T_{ijt} = l_{it} \forall i \in I, t \in \mathcal{T}$$

and where T_{ijt} is a free optimization variable for all suppressed variables and equals $T_{ijt}^{\text{non-suppressed}}$ otherwise.

This extension is straightforward but does not add anything in terms of improved performance. The reason is that the different years are effectively disconnected in this model, and maximum entropy is obtained by finding maximum entropy for each year separately.

To improve performance, we need to modify the objective function to consider the time development. A central idea is that when a value flips from zero to a public value or from a public value to zero, it is likely that the suppressed value is close to the cut-off. Such cases involve a small number of agents, and the effect is typically triggered when one agent randomly changes status. It seems less likely that several agents change their status simultaneously. In the next section, we discuss some ways of implementing time series information of this sort.

4.5.1 Incorporating Time Series Information

We first choose and fix the entry (i, j) in the trip distribution matrices, and consider what happens to the flow T_{ijt} for the years $t \in \mathcal{T}$. For some years t, it may happen that T_{ijt} is suppressed, while at other years, this flow might not be suppressed. We define a new reward parameter f_{ij} as follows:

$$f_{ij} = \frac{\#\text{ of years where } T_{ijt} \text{ is not suppressed}}{\#\text{ years in total}}$$

The intuition behind this parameter is as follows. Let h represent the highest suppressed value. The more years we have unsuppressed data for the entry (i, j), the more we believe that suppressed data in the other years are close to h — and most likely equal to h. We have implemented this in our objective function via the term:

$$\sum_{(i,j,t)\in\mathcal{V}} \Big(\exp\left[f_{ij}(h - T_{ijt})\right] - 1 \Big) \tag{4.2}$$

The larger the difference between h and T_{ijt}, the more we penalize the objective function, and the penalty is higher the larger the value on f_{ij}. This favors values of T_{ijt} that are close to h.

An alternative way of handling this is through a transition-based idea: If a suppressed value is either preceded by an non-suppressed value, superseded by an non-suppressed value, or both, a reasonable hypothesis is that the suppressed value is close to h. Let $\mathcal{D}$ denote the set of indices where one of the above-quoted properties holds. We can then consider a penalty on the form:

$$\sum_{(i,j,t)\in\mathcal{D}} \Big(\exp\left[(h - T_{ijt})\right] - 1 \Big) \tag{4.3}$$

Again, this term favors values of T_{ijt} that are close to h, but only if the variable in entry (i, j) at time t is super-seeded or presided by unsuppressed values. We tried several other ways of implementing penalties, but these proved unsuccessful in improving the accuracy in our training data.

4.5.2 Combining Time Series and Entropy

When we modify the objective function to include the penalty terms, the question of how they should be weighted arises. We consider the following objective function:

$$\begin{aligned} \max_{T_{ijt}} \quad & -\gamma \sum_{(i,j,t)\in\mathcal{V}} T_{ijt} \cdot \ln(T_{ijt}) - (1-\gamma)\delta \sum_{(i,j,t)\in\mathcal{V}} \Big(\exp\left[f_{ij}(h - T_{ijt})\right] - 1 \Big) \\ & \qquad - (1-\gamma)(1-\delta) \sum_{(i,j,t)\in\mathcal{D}} \Big(\exp\left[(h - T_{ijt})\right] - 1 \Big) \end{aligned} \tag{4.4}$$

Here $\gamma \in [0, 1]$ and $\delta \in [0, 1]$ are hyperparameters determining how much weight we should put on the different terms in the objective function. The case $\gamma = 1$ corresponds to the original formulation where all weight is put on entropy. If $\delta = 1$, the transition term has no impact, while the case $\delta = 0$ attributes full impact to the transition term.

We experimented with many different values of the hyperparameters on a different synthetic data set, our training data, with the same set of locations and number of years. The results indicated that a cluster of the consistently best results, in terms of accuracy, formed at, and close to, the hyperparameters $(\gamma, \delta) = (0.8, 0.2)$. Hence, $(\gamma, \delta) = (0.8, 0.2)$ are our hyperparameters of choice, giving the entropy, fraction, and transition terms an 80%, 4%, and 16% weight, respectively. However, the accuracy deviated only moderately across alternatives within this cluster, indicating that the choice of hyperparameters is not prone to overfit the model.

4.5.3 Results Based on Incorporating Information on Time Series

The results in Table 4.5.1, on the effect of incorporating information from time series and distances, reinforces the conclusion that entropy maximization performs well in disclosing information that was suppressed. The accuracy and SRMSE improvements are not substantial when either fractions in (4.2) or transitions in (4.3) are incorporated into the formulation of the maximization problem, with $\delta = 1$ and $\delta = 0$, respectively. It follows from Table 4.5.1 that adjusting for the effect of time series varies somewhat over the years. For the overall 20-year period, however, they each contribute to a reduction in SRMSE from about 0.39 to about 0.37, while Accuracy increases from about 0.75 to about 0.77.

Table 4.5.1. Results from disclosing suppressed cell information by a pure entropy maximization approach, and by approaches in addition incorporating time series information. The first three columns represent the year, the number of suppressed cells, and the number of commuters in the suppressed cells.

			Pure		$(\gamma,\delta)=(0.8,1)$		$(\gamma,\delta)=(0.8,0)$		$(\gamma,\delta)=(0.8,0.2)$	
Year	$\|CELL\|$	$\sum$	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE
1	66	89	0.6515	0.4654	0.7576	0.3651	0.6515	0.4654	0.7273	0.3873
2	67	85	0.7313	0.4086	0.7313	0.4086	0.8806	0.2724	0.8806	0.2724
3	66	80	0.7121	0.4763	0.7879	0.3800	0.7879	0.3800	0.8182	0.3518
4	65	77	0.6923	0.4683	0.6923	0.4683	0.7231	0.4442	0.7231	0.4442
5	69	84	0.6812	0.4638	0.6812	0.4638	0.7101	0.4422	0.7681	0.3956
6	73	93	0.6712	0.4501	0.8082	0.3437	0.8082	0.3437	0.8082	0.3437
7	67	84	0.7313	0.4134	0.7313	0.4134	0.7313	0.4134	0.7313	0.4134
8	64	76	0.8125	0.3646	0.7812	0.3939	0.8125	0.3646	0.8125	0.3646
9	65	78	0.7538	0.4134	0.6923	0.4623	0.7846	0.3867	0.7846	0.3867
10	64	77	0.8125	0.3599	0.7812	0.3887	0.8125	0.3599	0.8125	0.3599
11	68	89	0.7941	0.3467	0.7059	0.4144	0.6765	0.4346	0.7059	0.4144
12	66	88	0.7879	0.3454	0.8788	0.2611	0.6970	0.4129	0.8788	0.2611
13	67	91	0.7612	0.3598	0.7612	0.3598	0.7612	0.3598	0.8209	0.3116
14	67	85	0.7463	0.4307	0.8209	0.3336	0.8060	0.3852	0.8209	0.3336
15	70	94	0.7714	0.3560	0.7714	0.3560	0.7714	0.3560	0.7714	0.3560
16	65	78	0.7538	0.4134	0.7538	0.4134	0.7846	0.3867	0.7846	0.3867
17	68	86	0.8235	0.3322	0.8235	0.3322	0.9412	0.1918	0.9412	0.1918
18	72	97	0.7778	0.3499	0.7500	0.3711	0.7500	0.3711	0.7778	0.3499
19	75	108	0.7867	0.3208	0.9467	0.1604	0.7600	0.3402	0.8667	0.2536
20	76	109	0.8158	0.2993	0.7632	0.3393	0.7895	0.3199	0.8421	0.2771
Total	1360	1748	0.7537	0.3913	0.7721	0.3715	0.7721	0.3750	0.8044	0.3441

In the approach where both fractions and transitions are accounted for, with $(\gamma,\delta) = (0.8, 0.2)$, the suppressed information is to a larger degree disclosed, with accuracy of about 0.80 and SRMSE of about 0.34. However, it remains to be seen whether this improved fit has a significant practical impact on estimation results and predictions when data are implemented in a spatial interaction model. An alternative way of presenting the degree of replication is in terms of the matrix in Table 4.5.2.

Table 4.5.2. The replication matrix corresponding to a procedure with entropy maximization and time series adjustments; $(\gamma, \delta) = (0.8, 0.2)$. Aggregated over all the years.

	0	1	2
0	26	14	0
1	54	719	119
2	0	79	349

Consider, for example, all the origin-destination combinations with just one observed commuter. The procedure correctly predicted 719 of these 892 combinations. Notice also that the prediction never deviates from the actual number by more than one commuter over the entire period.

4.6 Evaluating Extended Rules of Suppression

As pointed out by Abowd and Schmutte (2019), the availability of linearly dependent statistics—such as marginal constraints—represents a potential for reconstructing confidential variables. Hence, privacy concerns may call for some action from the data-releasing agency in cases where such information is available. The agency can choose to suppress information on the marginal sums of the commuting matrix. This option is discussed in Section 4.6.1, while Section 4.6.2 addresses the option of raising the cut-off suppression value.

4.6.1 What If Marginal Totals Are Suppressed?

One option is to suppress the row sums, representing the spatial residential location pattern of the workers. Another option is to suppress the column sums, that is, information on the number of jobs in each of the towns in the geography. Finally, it is, of course, also an option to suppress both row and column sums. The results in Table 4.6.1 are based on the situations where one worker is added to all cells with 0 or 1 commuter in the first year. The results from experiments with time series information are not included in the table, as they did not contribute to replicating the actual commuting matrices over the years.

Table 4.6.1. Year-by-year comparisons of disclosing suppressed information in cases where row and column sums are both KNOWN, ROW sums are suppressed, COLUMN sums are suppressed, and BOTH rows and columns sums are suppressed, respectively.

			KNOWN		ROW		COLUMN		BOTH	
Year	$\|CELL\|$	$\sum$	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE
1	66	89	0.7273	0.3873	0.6212	0.4830	0.5303	0.5773	0.5455	0.5477
2	67	85	0.8806	0.2724	0.5373	0.6090	0.6269	0.5096	0.5373	0.6531
3	66	80	0.8182	0.3518	0.5606	0.6736	0.5909	0.5562	0.5000	0.6581
4	65	77	0.7231	0.4442	0.4154	0.7404	0.6000	0.5923	0.4769	0.7550
5	69	84	0.7681	0.3956	0.4783	0.6851	0.6957	0.4845	0.4928	0.7400
6	73	93	0.8082	0.3437	0.5753	0.5357	0.6164	0.4861	0.4795	0.6496
7	67	84	0.7313	0.4134	0.5672	0.6007	0.6269	0.5682	0.5821	0.6609
8	64	76	0.8125	0.3646	0.4844	0.7293	0.5312	0.6316	0.4531	0.7443
9	65	78	0.7846	0.3867	0.5385	0.7161	0.5692	0.6027	0.4923	0.7596
10	64	77	0.8125	0.3599	0.5000	0.6892	0.5469	0.5877	0.5000	0.6892
11	68	89	0.7059	0.4144	0.5441	0.5860	0.6029	0.6005	0.5147	0.6419
12	66	88	0.8788	0.2611	0.6667	0.5383	0.6515	0.4707	0.5455	0.5539
13	67	91	0.8209	0.3116	0.5821	0.5245	0.6567	0.4587	0.5672	0.5088
14	67	85	0.8209	0.3336	0.6119	0.5447	0.6716	0.4517	0.5821	0.5615
15	70	94	0.7714	0.3560	0.5857	0.5487	0.6857	0.4710	0.6429	0.5629
16	65	78	0.7846	0.3867	0.6154	0.6027	0.6154	0.6027	0.5077	0.6856
17	68	86	0.9412	0.1918	0.6176	0.4889	0.6471	0.5252	0.6029	0.5252
18	72	97	0.7778	0.3499	0.6667	0.4285	0.6528	0.4629	0.6250	0.4791
19	75	108	0.8667	0.2536	0.6800	0.3928	0.6933	0.4089	0.7467	0.3761
20	76	109	0.8421	0.2771	0.5789	0.4524	0.6579	0.4524	0.6711	0.4232
	1360	1748	0.8044	0.3441	0.5728	0.5739	0.6250	0.5219	0.5559	0.6041

Table 4.6.1 refers to four different situations of available information. The KNOWN situation is where both row and column sums are known, while row and column sums are suppressed in situations ROWS and COLUMNS, respectively. In the BOTH situation, both information on row sums and column sums are suppressed. In all the cases in Table 4.6.1, assume that we still have reliable information on the total number of workers and jobs in the geography—that is, the sum of all rows and columns, respectively. If so, we have information on the total number of suppressed workers but not how they are distributed between cells, rows, and columns, in the case where both row and column sums are suppressed.

The situation BOTH corresponds to the following optimization problem:

$$\max_{T_{ijt}} - \sum_{(i,j,t)\in\mathcal{V}} T_{ijt}\log(T_{ijt}) \quad s.t. \quad \sum_{i,j\in I} T_{ijt} = \sum_{i,j\in I} T_{ijt}^{\text{non-suppressed}} \forall t \in \mathcal{T}$$

The situations ROWS and COLUMNS are defined similarly. The results are relatively encouraging from the position of the data releasing agency. It follows from Table 4.6.1 that suppressing information on both marginal totals in rows and columns leads to a reduction of the accuracy from a level of around 0.80 to a level of about 0.55, and SRMSE increases from about 0.34 to about 0.60. However, in considering these figures, remember that they refer to only the suppressed cells of the commuting matrix. This means that the deviations between the replicated and the actual total matrices will be relatively minor also in the case where row and column sums are suppressed. Hence, it seems reasonable to hypothesize that these deviations at least lead to a less severe bias in estimating parameters representing commuting behavior than in cases where the suppressed information is ignored.

Another interesting result from Table 4.6.1 is that the potential to disclose the suppressed cells is not substantially increased if only row or column sums are suppressed. This, in particular, applies for the case where row sums are suppressed, on average resulting in an accuracy of around 0.57. This means that the statistical agency is recommended to suppress row sums, while privacy is not harmfully reduced when data on column sums are released. Suppressing just column sums leads to an accuracy of around 0.63, and information on the spatial distribution of jobs may be helpful for several research projects.

Suppressing rows sums can be more challenging from a privacy concern than suppressing column sums. Technically, it can be seen from the matrix of commuting flows in Table 4.3.1 that all the rows have specific cells with some commuters below the cut-off value of 3. At the same time, this is the case for 11 of the 12 columns of the matrix. Hence, the information loss of suppression is somewhat lower for the columns. There are reasons to think this is not just a coincidence specific to this data set. The observation that one column has large numbers reflects the tendency of firms and jobs to cluster in, for example, a regional center, that is, town E in our geography. This is due to urbanization economies and Marshallian agglomeration forces, giving rise to economies of localization and resulting in a location pattern where jobs are spatially more concentrated than the residential location pattern.

4.6.2 Experimenting With the Cut-Off Suppression Value

Another action from the statistical agency may be to raise the cut-off value from suppressing cell information. Table 4.6.2 provides results from experimenting with the cut-off value. The accuracy is found to fall considerably with increasing the cut-off value above 3. Correspondingly, the SRMSE is found to increase. According to our experiments, however, there is a low difference in accuracy and SRMSE for cut-off values of 4 and 5. This reflects the nature of our data, representing a relatively small population with only 12 zones and a modest variation in the number of low-valued cells in the matrix of commuting flows. Also, keep in mind that these values refer to only the part of the commuting matrix that is suppressed. Hence, an approach resulting in a value of SRMSE around 0.65 represents a substantial improvement compared to a procedure where values are set to zero in all suppressed cells. Even with a cut-off value of 5 observations, the entropy-based approach of disclosing suppressed information may be removing a substantial source of bias in estimating parameters in a spatial interaction model.

Table 4.6.2. Results based on various cut-off values. $(\gamma, \delta)^s$ represents the problem variant when row sums are suppressed.

			$(\gamma, \delta)^s = (0.8, 0.2)$	
Suppression level	$\lvert CELL \rvert$	$\sum$	Acc.	SRMSE
<3	1360	1748	0.5728	0.5739
<4	1561	2351	0.4478	0.6958
<5	1653	2719	0.4459	0.6538

4.7 Implementing Distance Information

It makes intuitive sense that a short distance increases the probability that the suppressed value is close to h, all else equal. This is according to the frequently cited Tobler's first law of geography, which states that "everything is related to everything else, but near things are more connected than distant things" (Tobler, 1970). It also matters somewhat how distances are defined in the model. In our experiments, the following two alternatives were considered. First, we employ *globally scaled distances* in which all distances are scaled relative to the longest observed distance of all the (i, j)-combinations in the sample. Second, we employ *locally scaled distances*, all distances are scaled relative to the longest distance observed for a specific origin i.

If a town has a highly accessible location, with many job opportunities in surrounding areas, its inhabitants may not consider long-distance commuting. The labor market accessibility, reflecting the distances to relevant job opportunities, may vary systematically across the towns. Thus, the hypothesis is that locally scaled distances capture relevant characteristics of the spatial structure and potentially contribute to explaining both the observed commuting pattern in the region and the spatial pattern of suppressed information. This hypothesis was supported in our experiments—locally scaled distances outperform globally scaled distances in replicating the actual commuting matrix.

If the distance between two towns is considerable, it appears more likely that the suppressed value is small than when the distance is small. As an example of how this could be implemented, we suggest the following—in which d_{ij} denotes the geographical distance and normalized to one:

$$\begin{aligned}\max_{T_{ijt}} \quad -\gamma \sum_{(i,j,t)\in\mathcal{V}} (d_{ij}+\epsilon)\cdot T_{ijt}\cdot \ln(T_{ijt}) - (1-\gamma)\delta \sum_{(i,j,t)\in\mathcal{V}} \Big(\exp\left[f_{ij}(h-T_{ijt})\right]-1\Big) \\ -(1-\gamma)(1-\delta) \sum_{(i,j,t)\in\mathcal{D}} \Big(\exp\left[h-T_{ijt}\right]-1\Big)\end{aligned} \tag{4.5}$$

4.7.1 Information When Marginal Totals Are Provided

Introducing locally scaled distances in the transition term adds nothing to disclose the suppressed data. The results are more or less identical to those corresponding to the approach where fractions and transitions are accounted for, with $(\gamma, \delta) = (0.8, 0.2)$, with no care taken to distances. We have experimented with several other ways of implementing distances into the objective function than what is represented by (4.5)—in which the distance parameter is added to the entropy term. However, the conclusion remains the same—none of these implementations led to significantly better performance.

Still, this does, of course, not on a general basis mean that distances should not be used in disclosing suppressed data in a situation where the marginal totals are known. Such an approach is based on an appealing and theoretically sound hypothesis. Our results may reflect that data are generated from a small geography with a low variation of values representing long-distance commuting. We cannot rule out the possibility that adjusting for distances contributes substantially to disclosing suppressed information in other test regimes, for example, based on an actual data set from an extensive system of towns and cities. Further exploration of these issues is left for future research.

4.7.2 Information When Marginal Totals Are Suppressed

The evaluation of adjusting for distances changes substantially in cases where marginal totals are suppressed. Table 4.7.1 provides results for a case where row sums are suppressed. In comparing these results to the last two columns of Table 4.6.1, it follows that distances employ a noticeable impact in disclosing the correct number of workers in the suppressed cells. The accuracy increases from around 0.56 to around 0.71 due to distance adjustments. Correspondingly, the SRMSE is reduced from around 0.60 to around 0.43. While suppressing the marginal sums decreases the accuracy efficiently, 60.6% of this reduction is recouped by implementing a local distance parameter to the objective function's entropy term. This supports the reasonable hypothesis that distances generate more order and better fit in cases with sparse information on suppressed information patterns.

Table 4.7.1. Accounting for locally scaled distances in the entropy term with suppressed row sums—$(\gamma, \delta)^{s,d} = (0.8, 0.2)$.

Year	1	2	3	4	5	6	7	8	9	10	
$\|CELL\|$	66	67	66	65	69	73	67	64	65	64	
$\sum$	89	85	80	77	84	93	84	76	78	77	
Acc.	0.5909	0.6716	0.7727	0.6769	0.7246	0.6712	0.7313	0.7344	0.7077	0.7500	
SRMSE	0.5000	0.4517	0.4308	0.5129	0.4638	0.4501	0.4134	0.4708	0.4848	0.4156	
Year	11	12	13	14	15	16	17	18	19	20	**Total**
$\|CELL\|$	68	66	67	67	70	65	68	72	75	76	**1360**
$\sum$	89	88	91	85	94	78	86	97	108	109	**1748**
Acc.	0.7059	0.6818	0.6418	0.7015	0.7143	0.7385	0.7941	0.7500	0.7333	0.7632	**0.7132**
SRMSE	0.4144	0.4523	0.4407	0.4307	0.4538	0.4623	0.3588	0.3711	0.3586	0.3393	**0.4324**

As demonstrated in Section 4.5.3, incorporating time series information improves disclosing suppressed information in cases where the marginal sums are known. The results in Section 4.6.1, on the other hand, demonstrated that time series information did not significantly add to the performance in cases where marginal totals are suppressed. For distances, the results are the opposite. Information on distances adds substantially to the potential of disclosing suppressed information in instances where row and/or column sums are not known. We hypothesize that the hard marginal total constraints limit the feasible space to such an extent that the soft constraints from distance information do not provide any added value beyond what the time series considerations provide.

The discussion in Section 4.6.2 demonstrated that raising the cut-off value of cell suppression represents an efficient action to reduce the accuracy and preserve privacy in providing data. This can, of course, also be done in cases where the marginal sums are suppressed. By comparing the results in Table 4.7.2 to the results in Table 4.6.2, it once again follows that adjusting for information on distances leads to a considerable increase in the accuracy and a corresponding reduction in the SRMSE.

Table 4.7.2. Results based on various cut-off values in a case with suppressed row and column sums.

			$(\gamma,\delta)^{s,d} = (0.8, 0.2)$	
Suppression level	$\|CELL\|$	$\sum$	Acc.	SRMSE
<3	1360	1748	0.7132	0.4324
<4	1561	2351	0.6028	0.4590
<5	1653	2719	0.5590	0.4662

Table 4.7.3. The replication matrix corresponding to a procedure with entropy maximization and distance adjustments in a case with a cut-off value of 5. Aggregated over all the years.

	0	1	2	3	4
0	4	23	6	0	0
1	60	579	127	11	0
2	13	193	260	123	25
3	3	17	74	61	47
4	0	0	1	6	20

The replication matrix in Table 4.7.3 refers to a cut-off value of 5. This matrix gives a clear perception that entropy-based allocation of suppressed information is superior to assigning a value of 0 in the relevant cells. Specifically, while the accuracy remains moderately high at 0.5590, only around 4.6% of the 1653 predictions have a deviation greater than one from the true values. Just 3 of the 1653 predictions deviate by more than two from the true values. This justifies a hypothesis that entropy-based allocation of suppressed information removes a potentially serious bias in estimating parameters representing spatial labor market mobility. This hypothesis will be examined in the section to follow.

4.8 Estimation Results Based on the Doubly Constrained Gravity Model

As pointed out in Section 4.4.1, the modified maximum entropy problem has as its solution a multinomial logit model. In (4.1), this model formulation was given by $T_{ij} = A_i B_j e^{-\beta c_{ij}}$, where $A_1, ..., A_n, B_1, ..., B_n$ are constants. Additionally, let c_{ij} be a measure of spatial separation between an origin i and a destination j, defined by the generalized cost of traveling from origin i to destination j. The doubly constrained gravity model is an alternative, equivalent formulation of this trip distribution model. In a commuting context, let O_i represent the number of commuters originating from town i, while D_j is the observed number of trips with destination in zone j. A_i and B_j are the balancing factors ensuring that $\sum_j T_{ij} = O_i$ and $\sum_i T_{ij} = D_j$. A standard formulation of a doubly constrained gravity model then is:

$$T_{ij} = A_i O_i B_j D_j e^{-\beta c_{ij}}, \quad A_i = \left[\sum_j B_j D_j e^{-\beta c_{ij}}\right]^{-1}, \quad B_j = \left[\sum_i A_i O_i e^{-\beta c_{ij}}\right]^{-1} \tag{4.6}$$

For more details on the theoretical foundation of this model, see for instance Erlander and Stewart (1990) or Sen and Smith (1995). O_i and D_j represent characteristics of the origin and destination towns expected to influence the volume of work-related trips for the specific combination of towns. The literature also provides a discussion of the need to account for spatial characteristics that are not captured by the balancing factors—see Fotheringham (1983), Pellegrini and Fotheringham (2002), Gitlesen and Thorsen (2000) and Persyn and Torfs (2016). Leaving out relevant characteristics of spatial structure introduces a source of estimation bias of spatial interaction parameters—see Tiefelsdorf (2003). This article considers potential bias resulting from suppressed information on commuting. First, we estimated the distance deterrence parameter β following from the standard doubly constrained gravity model in a case where no data are suppressed, that is, for the true matrix of commuting trips. These estimates are represented by the red curve in Figure 4.8.1. The standard errors of the parameter estimates are calculated by bootstrapping, and they are also included in Figure 4.8.1.

The green curve in Figure 4.8.1 represents estimates of the distance deterrence parameter in the case with suppressed information, with a value of 0 in cells with h commuters or fewer. The estimates do not deviate much from the estimates based on the non-suppressed matrix of commuting flows. However, as illustrated in the figure, there is a systematic difference over the years. Ignoring the suppressed information leads to consistently higher estimates of the distance deterrence parameter. Carpenter et al. (2022) claim that bias may follow from the reasonable possibility that less populous geographies have a higher share of suppressed cells—similarly, in commuting, suppressed cells may mainly come about for long-distance journeys to work.

The differences between the estimates from the non-suppressed and the suppressed matrices of commuting trips are not significantly different from 0—the confidence interval of all the estimated differences incorporates 0. When evaluating our approach's performance, keep in mind that there are two different sources of uncertainty in these matrices. One source is purely random and is caused by the actions of a finite number of agents. The other source is systematic and is caused by the suppression of data. Suppression of data will, in most cases, lead to a lower number of long-distance commuters, leading to systematically lower estimates for the distance deterrence parameter. A central finding in this article is that our methods reduce the systematic bias to such small numbers that it is effectively zero.

When the total population increases, the uncertainty due to random choices goes down. In an artificial experiment, we split all the zones into two sub-zones. In the corresponding 24×24 trip-distribution, we allocated the numbers in the original 12×12 trip-distribution to each 2×2 block. The total number of agents in the system increased by four. As expected, the standard deviation in the parameter estimates was roughly 50% of the original values. The optimal parameter values changed, but when we compared the non-suppressed and suppressed matrices, the systematic bias slightly increased. This hints that the systematic bias may dominate the random bias in a system with many agents distributed across a geography with many zones.

We have also estimated the distance deterrence parameter from commuting matrices where the suppressed information is disclosed by the methods explained in the previous sections, in the case where the marginal totals are known. In this case, the estimates almost precisely match the corresponding estimates for the non-suppressed trip distribution, with no differences in the first three decimals after rounding. To a slightly lower degree, this is also the case for the data matrix where row sums are suppressed. The suppressed information is next disclosed by the procedure where the entropy-based method is supplemented by locally scaled distances. The results based on this case are illustrated in Figure 4.8.1, represented by the blue curve. The curves marked in red and blue overlap for the major parts and can only be distinguished from one another in a few segments. Hence, from a practical point of view, the bias resulting from suppressed information is also removed in this case.

We have also experimented by varying the cut-off value for suppressing information in the cells of the OD-matrix. As expected, higher cut-off values result in higher estimates of the distance deterrence parameter. However, without entering into details, the increase is not found to be very sensitive to the chosen cut-off value. The reason for this can be found by considering the geography and the synthetic population underlying the experiments. Due to the considerable computing time involved, we kept our agent-based population relatively small, distributed across only 12 towns. As illustrated in Figure 4.3.1 on page 113, the towns are organized into three clusters. The distances are long between the three clusters but short between the towns within each cluster. This explains a pattern of commuting flows with just a few commuters between towns belonging to different clusters but considerable commuting between towns within the same cluster. See the matrix of commuting flows in Table 4.3.1. In many, more densely populated geographies, a larger continuum can be expected in both sizes and distances across towns and cities. The commuting pattern will reflect more heterogeneity in sizes and distances, with more continuous variation expected in the number of commuters between the different origin-destination combinations. In such a case, the $\hat{\beta}$'s are expected to be considerably more sensitive to the cut-off value of cell suppression. Hence, cell suppression represents a source of seriously biased estimates and predictions. However, entropy-based approaches to disclosing suppressed information may eliminate this source of bias.

Our experiments are run for a relatively sparsely populated geography. It is a reasonable hypothesis that the estimation bias resulting from suppression is more severe and significant for real geographies with more towns, more workers, and more substantial variation in distances, particularly between towns far from each other. We leave the testing of such a hypothesis for future research.

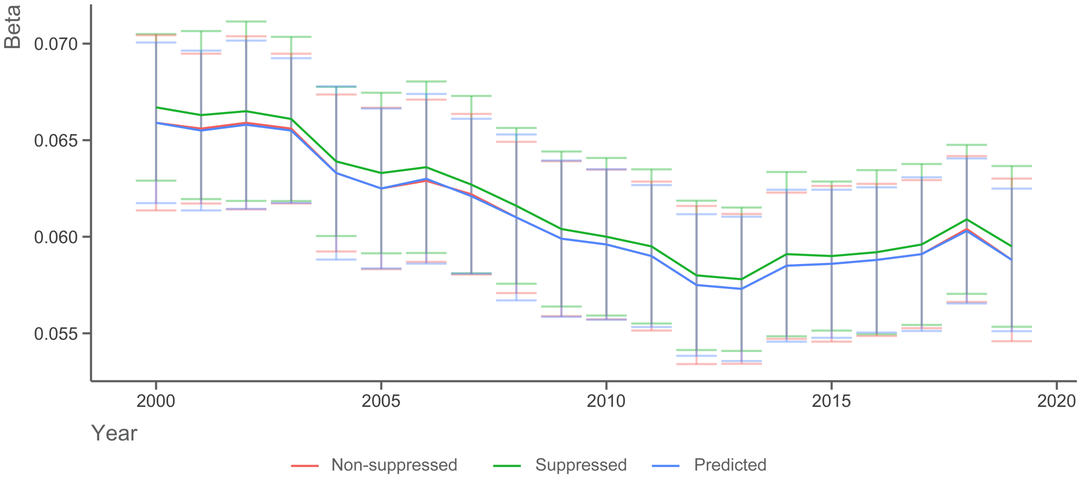

Figure 4.8.1. $\hat{\beta}$ values for non-suppressed, suppressed, and predicted $\big((\gamma,\delta)^{s,d} = (0.8, 0.2)\big)$ data sets.

4.9 Summary and Concluding Remarks

There has been a growing concern for privacy issues recently, reinforced by the increasing access to micro-data. This has been inducing statistical agencies to introduce statistical disclosure limitations in releasing data. The most commonly used method for tabular data is cell suppression.

This represents a challenge for empirical research (Abowd & Schmutte, 2019), representing a source of measurement error that may lead to biased parameter estimates and/or increased standard errors (Carpenter et al., 2022). In this article, we discuss how this utility loss in research can be reduced through attempts to disclose the information represented by the suppressed cells.

Journeys-to-work may not in itself represent a very sensitive type of information. However, it has the potential of being linked to external data sources that can be subject to malicious use by intruders. Our discussion is conducted in terms of a case where the spatial dimension leads to cells in an origin-destination matrix with just a few commuters, facilitating identity disclosures. In addition, our results are generally valid for other, more sensitive, social matters than commuting.

Based on synthetic data mirroring the Norwegian population, we demonstrate that a constrained entropy maximizing approach to a large degree succeeds in disclosing the information hidden by suppressing information in cells with less than three commuters. The constraints represent the marginal sums in the matrix of commuting trips, defining the given and known number of workers living and working in each zone of the geography. We also demonstrate that incorporating information on time series in an entropy-maximizing approach adds significantly to disclosing suppressed information. In contrast, data on distances does not turn out to be relevant in the case with a cut-off suppression value of 3.

As pointed out by Abowd and Schmutte (2016), statistical limitation disclosure is ignorable if correct inferences are made without explicitly accounting for SDL, but Carpenter et al. (2022) in general call for greater awareness in documenting the methods used in estimation based on suppressed cell data sets. We provide estimates of the distance deterrence parameter in a standard doubly constrained gravity model for data sets based on different treatments of suppressed cells. Ignoring the information in the suppressed cells is found to cause a bias, as expected, with an overvalued estimate of the distance deterrence in commuting. This bias is largely eliminated when entropy-based approaches disclose the suppressed information. Quantitatively, these effects are not found to be strong for our synthetic data, but ignoring suppressed information is demonstrated to be potentially harmful, and the results are encouraging because they prove potential for entropy-maximizing approaches to avoid biased estimates and predictions.

It is well known in the literature that if many linearly dependent statistics are available, there is a substantial potential for a reconstruction attack, a reconstruction of confidential variables, which is a data breach—see Abowd and Schmutte, 2019. The marginal sums in the matrix of commuting flows are examples of this kind of statistics. What if statistical agencies undertake a secondary suppression to avoid such attacks, that is, to suppress information on the marginal sums? According to our results, leaving out information on marginal totals does not lead to a substantial limitation of the possibility of disclosing suppressed information about commuting trips. An alternative step to preserve privacy is that the statistical agency chooses to increase the cut-off value of cell suppression. Still, our results demonstrate that an entropy-based procedure can remove a substantial source of bias in cases with a higher cut-off value of cell suppression.

Another problem we have addressed is the potential of integrating additional information on the system into the entropy maximizing procedure, discussing the degree to which this contributes to disclosing the suppressed information. In many cases, information on commuting trips is available for several successive periods, like the 20 years period that we consider. In addition, information is generally available on distances between alternative origin-destination combinations of journeys to work. We find that information on time series is useful, while data on distance are not in a case with known marginal totals. This conclusion is reversed when we have no information on marginal totals. Distances contribute to significantly better fit in such a case with sparse information on the pattern of suppressed information, while time series information does no longer contribute. Based on our synthetic data experiments, we also found that the potential of an identity disclosure is not substantially increased if only row or column sums are suppressed.

The waterfall chart in Figure 4.9.1 summarizes the main results of different approaches to disclosing suppressed information. First, a pure entropy maximization provides a 75.37% prediction accuracy that follows from suppressing all cells with less than three observations. Second, accounting for time series information provides another 5.07 percentage points of the prediction accuracy. Third and fourth, suppressing row sums reduces the accuracy by 23.16 percentage points, while adding locally scaled distances on the entropy term recoups some of this loss by 14.04 percentage points. Fifth, this accuracy increase is lost when the cells are suppressed at a threshold of 5 instead of 3, with the accuracy falling by 15.42 percentage points. Moreover, despite the fact that prediction inaccuracies exists, the predictions largely removes the estimation bias of the distance deterrence parameter in a doubly constrained gravity model.

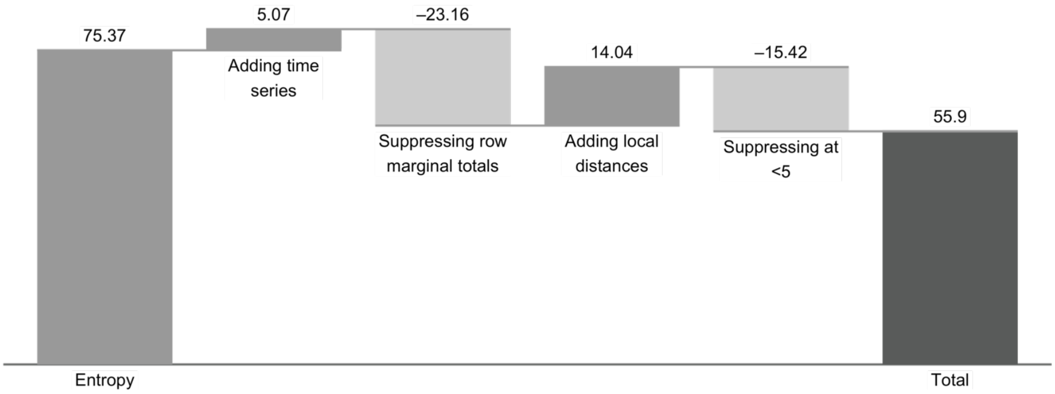

Figure 4.9.1. While marginal total suppression and increasing the cut-off value decreases the prediction accuracy, incorporating distances recoups 60% of the suppression loss.

The literature distinguishes between identity disclosure and inferential disclosure, where the former can be argued to be the most important kind of disclosure to avoid (G. Duncan and Lambert, 1989, Airoldi et al., 2011, and Abowd and Schmutte, 2016). As stated in the introduction, Abowd and Schmutte (2016) claimed that it is, from a probabilistic perspective, impossible to release data without compromising confidentiality. They also claim that there should be less trepidation in using data that have been the subject of statistical disclosure limitation—in general, it is not more serious than other sources of non-ignorable missing data.

Abowd and Schmutte (2019) recommended that adding noise to the micro-data may be a better alternative to suppression, and Abowd and Schmutte (2016) claimed that data-releasing agencies are becoming more open to the use of noise-inferred methods in producing data tables. This is useful in reaching unbiased parameters, as opposed to what is in general following from approaches with count-based suppression. On the other hand, estimates based on noise infusion are less precise, which according to Chetty and Friedman (2019) is a crucial drawback and a primary concern in much research. As a general recommendation, statistical agencies should be encouraged to use methods that, to a low degree, limit the statistical validity of the study and, to a high degree, makes it possible to reach results that correspond to the results that would follow from non-suppressed data.

The main contribution of this article is twofold. First, we demonstrate that suppressing the information in cells with a low number of observations does not necessarily preserve privacy adequately. The agencies should consider suppressing marginal totals, using higher cut-off values, and/or other methods, such as noise infusion. Second, our results strongly recommend that researchers develop sound methods to adjust for suppressed information rather than just ignoring the cells by setting the values equal to zero. Well-founded methods to adjust for suppressed information can potentially remove a source of harmful bias in estimating basic parameters and prevent invalid predictions, for instance, on important policy issues. Our study is performed from a spatial labor market interaction perspective, but it is at least a reasonable hypothesis that these conclusions are valid also from a more general perspective.

References

Abowd, J. M., & Schmutte, I. M. (2016). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, *2015*(1), 221–293.

Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, *109*(1), 171–202.

Airoldi, E. M., Bai, X., & Malin, B. A. (2011). An entropy approach to disclosure risk assessment: Lessons from real applications and simulated domains. *Decision Support Systems*, *51*(1), 10–20.

Anas, A. (1983a). Cities and complexity: Understanding cities through cellular automata, agent based models and fractals. *Transportation Research Part B: Methodological*, *17*(1), 13–23.

Anas, A. (1983b). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, *17*(1), 13–23.

Bartik, T. J., Biddle, S. C., Hershbein, B. J., & Sotherland, N. D. (2018). Wholedata: Unsuppressed county business patterns data: Version 1.0 [dataset]. *Kalamazoo: WE Upjohn Institute for Employment Research.*

Bezzi, M. (2007). An entropy based method for measuring anonymity. *2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops-SecureComm 2007*, 28–32.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 200–217.

Carpenter, C. W., Van Sandt, A., & Loveridge, S. (2022). Measurement error in us regional economic data. *Journal of Regional Science*, *62*(1), 57–80.

Chetty, R., & Friedman, J. N. (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples. *AEA Papers and Proceedings*, *109*, 414–20.

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, *7*(2), 207–217.

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). Why statistical confidentiality? In *Statistical confidentiality* (pp. 1–26). Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284.

Erlander, S. (2010). *Cost-minimizing choice behavior in transportation planning: A theoretical framework for logit models*. Springer Science & Business Media.

Erlander, S., & Stewart, N. F. (1990). *The gravity model in transportation analysis: Theory and extensions* (Vol. 3). Vsp.

Fotheringham, A. S. (1983). A new set of spatial-interaction models: The theory of competing destinations. *Environment and Planning A: Economy and Space*, *15*(1), 15–36.

Gholami, A., Thorsen, I., & Ubøe, J. (2022). *An agent-based approach to study spatial structure effects on estimated distance deterrence in commuting*. NHH working paper.

Gitlesen, J. P., & Thorsen, I. (2000). A competing destinations approach to modeling commuting flows: A theoretical interpretation and an empirical application of the model. *Environment and Planning A*, *32*(11), 2057–2074.

Gjestland, A., Thorsen, I., & Ubøe, J. (2006). Some aspects of the intraregional spatial distribution of local sector activities. *The Annals of Regional Science*, *40*(3), 559–582.

González-Vidal, A., Rathore, P., Rao, A. S., Mendoza-Bernal, J., Palaniswami, M., & Skarmeta-Gómez, A. F. (2020). Missing data imputation with bayesian maximum entropy for internet of things applications. *IEEE Internet of Things Journal*, *8*(21), 16108–16120.

Guldmann, J.-M. (2013). Analytical strategies for estimating suppressed and missing data in large regional and local employment, population, and transportation databases. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, *3*(4), 280–289.

Irwin, E. G. (2010). New directions for urban economic models of land use change: Incorporating spatial dynamics and heterogeneity. *Journal of Regional Science*, *50*(1), 65–91.

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, *110*(9), 2819–58.

Li, C., Miklau, G., Hay, M., McGregor, A., & Rastogi, V. (2015). The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB Journal*, *24*(6), 757–781.

Majeed, A., & Lee, S. (2020). Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data. *Applied Intelligence*, *50*(8), 2555–2574.

Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, *105*(1), 272–98.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1–29.

Mattson, L.-G., & Weibull, J. (2002). Probabilistic choice and procedurally rationality. *Games and Economic Behavior*, *41*(1), 61–78.

McArthur, D. P., Thorsen, I., & Ubøe, J. (2012). Labour market effects in assessing the costs and benefits of road pricing. *Transportation Research Part A: Policy and Practice*, *46*(2), 310–321.

McArthur, D. P., Thorsen, I., & Ubøe, J. (2010). A micro-simulation approach to modelling spatial unemployment disparities. *Growth and Change*, *41*(3), 374–402.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, *3*(4), 303–328.

Page, S. E. (1999). On the emergence of cities. *Journal of Urban Economics*, *45*(1), 184–208.

Pellegrini, P. A., & Fotheringham, A. S. (2002). Modelling spatial choice: A review and synthesis in a migration context. *Progress in Human Geography*, *26*(4), 487–510.

Persyn, D., & Torfs, W. (2016). A gravity equation for commuting with an application to estimating regional border effects in belgium. *Journal of Economic Geography*, *16*(1), 155–175.

Rajasekaran, S., Harel, O., Zuba, M., Matthews, G., & Aseltine, R. (2009). Responsible data releases. *Industrial Conference on Data Mining*, 388–400.

Rodrigues, J. D. (2016). Maximum-entropy prior uncertainty and correlation of statistical economic data. *Journal of Business & Economic Statistics*, *34*(3), 357–367.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Salazar-Gonzalez, J.-J. (2004). Mathematical models for applying cell suppression methodology in statistical data protection. *European Journal of Operational Research*, *154*(3), 740–754.

Sen, A., & Smith, T. E. (1995). Gravity models: An overview. *Gravity Models of Spatial Interaction Behavior*, 49–152.

Shlomo, N. (2018). Statistical disclosure limitation: New directions and challenges. *Journal of Privacy and Confidentiality*, *8*(1).

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Tiefelsdorf, M. (2003). Misspecifications in interaction model distance decay relations: A spatial structure effect. *Journal of Geographical Systems*, *5*(1), 25–50.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*(sup1), 234–240.

Train, K. (2003). *Discrete choice methods with simulation.* Cambridge University Press.

Willenborg, L., & De Waal, T. (2012). *Elements of statistical disclosure control* (Vol. 155). Springer Science & Business Media.

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*(3), 253–269.

Wilson, A. (2010). Entropy in urban and regional modelling: Retrospect and prospect. *Geographical Analysis*, *42*(4), 364–394.

4.A Mathematical Programming Models

This appendix provides a complete overview of the models and model variants presented in previous sections. Two models were employed—first, a non-linear model to predict the number of commuters from an entropy perspective and possibly from a time series and distance perspective. The second model is an integer model that fits the predictions to integer numbers while respecting the necessary constraints.

The first model has five problem variants, while the second has two. The first model's first objective function is a pure entropy function. Second, the objective function is expanded with time series considerations. Third, marginal totals are suppressed row-wise while keeping the second objective function. Fourth, the objective function is expanded with locally scaled distances on the entropy term. Fifth, the parameter h denoting the highest suppressed value is increased from two to five. The same objective function is applied across both variants for the second model, but they differ on whether to suppress row-wise marginal totals.

Calculations were done on a Linux Ubuntu 18.04 virtual machine with an Intel Xeon Platinum 8272CL hyper-threaded processor with 18 cores and 2.6 GHz clock speed. The mathematical programming environment AMPL and the commercial optimization solvers BARON and CPLEX were used to run the optimization problems.

Table 4.A.1. Model components.

Sets	
I	An ordered set of locations.
$\mathcal{T}$	An ordered set of years.
Subsets	
$\mathcal{C} \subset (I, I, \mathcal{T})$	The set of constants.
$\mathcal{V} \subset (I, I, \mathcal{T})$	The set of decision variables—i.e. the elements that are suppressed at time t.
$\mathcal{D} \subset (I, I, \mathcal{T})$	The set of suppressed (i, j, t) elements that is preceded by a non-suppressed value, superseded by a non-suppressed value, or both.
Variables	
$T_{ijt} \geq 0$	The predicted number of commuters.
$x^{+}_{ijt} \geq 0$	Positive deviation of the difference between the decimal prediction and the integer-imposed prediction.
$x^{-}_{ijt} \geq 0$	Negative deviation of the difference between the decimal prediction and the integer-imposed prediction.
Parameters	
$\epsilon \geq 0$	A minuscule small amount.
$f_{ijt} \geq 0$	Fraction of unsuppressed data across years for location (i, j).
$\gamma \geq 0$	Objective function hyperparameter weight of the entropy term relative to the time series terms.
$\delta \geq 0$	Objective function hyperparameter weight of the fraction term relative to the transition term.
$d_{ij} \geq 0$	Locally scaled distance parameter from location i to j.
$l_{it} \geq 0$	Number of workers living in location i at time t.
$e_{jt} \geq 0$	Location j's employment opportunities at time t.
$T^{non-suppressed}_{ijt} \geq 0$	Number of workers commuting from living location i to employment location j at time t.
$h \geq 0$	The highest suppressed value.
$T^{*}_{ijt} \geq 0$	The non-linear model's solution of T_{ijt}, which is a parameter in the integer model.

4.A.1 Constraints

This section covers eleven constraints that have been used in at least some of the five problem variants of the first model and the two different problem variants of the second model. Which constraints have been used in which problem variant can be seen from Table 4.A.2.

Table 4.A.2. Constraints used across problem variants. Y represents "Yes"—an inclusion of the constraint—while the no value N represents an omission of the constraint in the problem variant.

	Entropy		Time Series		Suppressing Row Sums		Local Distances		Suppressing <5	
Model	1	2	1	2	1	2	1	2	1	2
(4.7)–(4.8)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
(4.9)–(4.10)	Y	Y	Y	Y	N	N	N	N	N	N
(4.11)	N	N	N	N	Y	Y	Y	Y	Y	Y
(4.12)	N	Y	N	Y	N	Y	N	Y	N	Y
(4.13)–(4.14)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
(4.15–4.17)	N	Y	N	Y	N	Y	N	Y	N	Y

Employment opportunities

$$\sum_{(i,j,t)\in\mathcal{V}} T_{ijt} + \sum_{(i,j,t)\in\mathcal{C}} T_{ijt}^{non-suppressed} \geq e_{jt} - \epsilon \quad \forall j \in I, t \in \mathcal{T} \tag{4.7}$$

$$\sum_{(i,j,t)\in\mathcal{V}} T_{ijt} + \sum_{(i,j,t)\in\mathcal{C}} T_{ijt}^{non-suppressed} \leq e_{jt} + \epsilon \quad \forall j \in I, t \in \mathcal{T} \tag{4.8}$$

Constraints (4.7) and (4.8) state that the total number of predicted commuters and already known commuters who are commuting to location j in year t must equal the number of job opportunities at location j in year t. While it would generally be sufficient with one equality constraint to ensure this, the BARON solver's progress was substantially faster when allowing the left-hand side of the constraints to lie within the small interval of $[e_{jt} - \epsilon; e_{jt} + \epsilon]$ instead of it to be exactly equal to e_{jt}. Since the second model fits the predictions to integer values, deviating marginally from an integer total made no practical difference in the first model.

Residents

$$\sum_{(i,j,t)\in\mathcal{V}} T_{ijt} + \sum_{(i,j,t)\in\mathcal{C}} T_{ijt}^{non-suppressed} \geq l_{it} - \epsilon \quad \forall i \in I, t \in \mathcal{T} \tag{4.9}$$

$$\sum_{(i,j,t)\in\mathcal{V}} T_{ijt} + \sum_{(i,j,t)\in\mathcal{C}} T_{ijt}^{non-suppressed} \leq l_{it} + \epsilon \quad \forall i \in I, t \in \mathcal{T} \tag{4.10}$$

$$\sum_{(i,j,t)\in\mathcal{V}} T_{ijt} + \sum_{(i,j,t)\in\mathcal{C}} T_{ijt}^{non-suppressed} = \sum_{i\in I} l_{it} \quad \forall t \in \mathcal{T} \tag{4.11}$$

Similarly, constraints (4.9) and (4.10) states that the total number of predicted commuters and already known commuters who are commuting from location i in year t must equal the number of workers living in location i in year t. These constraints are employed for the first two problem variants of the first model, in which the objective function is either a pure entropy function or expanded with time series considerations. Constraint (4.11), however, is the suppressed marginal total constraint and states that the total number of predicted commuters and already known commuters in year t must equal the number of workers living in the system in year t.

Deviation relation

$$T^{*}_{ijt} - T_{ijt} = x^{+}_{ijt} - x^{-}_{ijt} \quad \forall (i,j,t) \in \mathcal{V} \tag{4.12}$$

The first model provides a decimal solution to the number of commuters, while the second model fits this solution to integers by minimizing the total absolute deviation from the first model. The second model employs constraint (4.12) to do so. The left-hand side of the constraint is the deviation for each (i,j,t), and the right-hand side has one non-negative decision variable for positive deviation and one non-negative decision variable for negative deviation from the predicted value in the first model. The total of these two decision variables makes up the absolute deviation for each (i,j,t).

Variable constraints

$$T_{ijt} \geq 0 \quad \forall (i,j,t) \in \mathcal{V} \tag{4.13}$$

$$T_{ijt} \leq h \quad \forall (i,j,t) \in \mathcal{V} \tag{4.14}$$

$$T_{ijt} \in \mathbb{N}_0 \quad \forall (i,j,t) \in \mathcal{V} \tag{4.15}$$

$$x^{+}_{ijt} \geq 0 \quad \forall (i,j,t) \in \mathcal{V} \tag{4.16}$$

$$x^{-}_{ijt} \geq 0 \quad \forall (i,j,t) \in \mathcal{V} \tag{4.17}$$

Finally, there are the variable constraints. For both models, the predicted number of commuters T_{ijt} must be non-negative (constraint (4.13)) and less than or equal to the highest number of suppressed commuters, h (constraint (4.14)) for all $(i,j,t) \in \mathcal{V}$. The second model enforces stricter requirements to T_{ijt} by also enforcing it to be an integer in constraint (4.15). The second model also enforces non-negative deviation variables for both the positive deviation (constraint (4.16)) and the negative deviation (constraint (4.17)).

4.A.2 Objective Function

Five problem variants have been presented in this paper—namely, a pure entropy model, a variant that also incorporates time series, a variant that suppresses the marginal row sums, a variant that adds local distances to the entropy term, and a variant that suppresses cells at a higher threshold of < 5 instead of < 3.

Table 4.A.3. Objective functions used across problem variants.

	Entropy		Time Series		Suppressing Row Sums		Local Distances		Suppressing <5	
Model	1	2	1	2	1	2	1	2	1	2
(4.18)	Y	N	N	N	N	N	N	N	N	N
(4.19)	N	N	Y	N	Y	N	N	N	N	N
(4.20)	N	N	N	N	N	N	Y	N	Y	N
(4.21)	N	Y	N	Y	N	Y	N	Y	N	Y

The pure entropy model's objective function in the first model is (4.18).

$$\min_{T_{ijt}} \sum_{(i,j,t)\in\mathcal{V}} T_{ijt} \cdot \ln(\epsilon + T_{ijt}) \tag{4.18}$$

When incorporating time series considerations into the first model, the objective function is expanded to (4.19). The same objective function is used for the problem variant that suppresses the marginal row sums.

$$\begin{aligned} \min_{T_{ijt}} \quad \gamma \sum_{(i,j,t)\in\mathcal{V}} T_{ijt} \cdot \ln(\epsilon + T_{ijt}) + \\ (1-\gamma)\delta \sum_{(i,j,t)\in\mathcal{V}} \left(\exp\left[f(h - T_{ijt}) \right] - 1 \right) + \\ (1-\gamma)(1-\delta) \sum_{(i,j,t)\in\mathcal{D}} \left(\exp[h - T_{ijt}] - 1 \right) \end{aligned} \tag{4.19}$$

For the first model's problem variants that add local distances to the model and later suppress cells at a higher threshold of < 5 instead of < 3, a local distance parameter is added to the entropy term of the objective function.

$$\begin{aligned}\min_{T_{ijt}} \quad & \gamma \sum_{(i,j,t)\in\mathcal{V}} (d_{ij}+\epsilon)\cdot T_{ijt}\cdot \ln(\epsilon + T_{ijt})+ \\ & (1-\gamma)\delta \sum_{(i,j,t)\in\mathcal{V}} \left(\exp\left[f(h-T_{ijt})\right]-1\right)+ \\ & (1-\gamma)(1-\delta) \sum_{(i,j,t)\in\mathcal{D}} \left(\exp[h-T_{ijt}]-1\right)\end{aligned} \tag{4.20}$$

Finally, the second model attempts to minimize the total absolute deviation that occurs when deciding the integer values of the predictions while respecting the constraints specified in Table 4.A.2 for this model across both its variants. This total is defined as:

$$\min_{x^+_{ijt},x^-_{ijt}} \sum_{(i,j,t)\in\mathcal{V}} (x^+_{ijt}+x^-_{ijt}) \tag{4.21}$$

That is why.