# Fleet management optimisation for ride-hailing services: From mixed traffic to fully automated environments

Fan, Q.

**DOI**
10.4233/uuid:eb3368a4-f45d-493f-8270-68d20a0309e3

**Publication date**
2025

**Document Version**
Final published version

**Citation (APA)**
Fan, Q. (2025). *Fleet management optimisation for ride-hailing services: From mixed traffic to fully automated environments*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:eb3368a4-f45d-493f-8270-68d20a0309e3

**Important note**
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

# FLEET MANAGEMENT OPTIMIZATION FOR RIDE-HAILING SERVICES

## FROM MIXED TRAFFIC TO FULLY AUTOMATED ENVIRONMENTS

Qiaochu FAN

# Fleet management optimisation for ride-hailing services: from mixed traffic to fully automated environments

Qiaochu Fan

# Fleet management optimisation for ride-hailing services: from mixed traffic to fully automated environments

**Dissertation**

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van den Hagen
chair of the Board for Doctorates
to be defended publicly on:
Tuesday, 29th April 2025 at 12:30

by

**Qiaochu Fan**
Master of Science in Transportation Planning and Management
Southwest Jiaotong University
born in Jinan, Shandong

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

| | |
|---|---|
| Rector Magnificus | Chairperson |
| Dr.ir. J. T. van Essen | Delft University of Technology, promotor |
| Dr.ir. G. Homem de Almeida Correia | Delft University of Technology, promotor |

Independent members:

| | |
|---|---|
| Prof.dr.ir. B. van Arem | Delft University of Technology |
| Prof.dr. K. An | Tongji University, China |
| Prof.dr.ir. K. I. Aardal | Delft University of Technology |
| Prof.dr. O. Cats | Delft University of Technology |
| Dr.ir. Y. Maknoon | Delft University of Technology, reserve member |

The research leading to this dissertation has received funding from the China Scholarship Council (CSC) and Delft University of Technology.
**TRAIL Thesis Series no. T2025/4, the Netherlands Research School TRAIL**

TRAIL
P.O. BOX 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-359-6

# Acknowledgement

Putting gratitude into words is a challenge because words can never fully capture the depth of my sincere appreciation. At first, I wanted to keep all my gratitude in my heart. But Gonçalo pointed out that gratitude should not work that way - it needs to be expressed. So, let me attempt to express it here, aiming to keep this acknowledgement concise enough to not exceed 100 pages.

First of all, I would like to thank my supervisors, Theresia and Gonçalo. Their guidance, patience and support throughout my PhD journey have been invaluable and I am truly grateful to have had them as my supervisors. Our meetings and discussions, where we brainstormed and exchanged ideas, were enriching. I have learnt so much from your knowledge and experience, which has made me a better researcher. Theresia, your ability to balance being a great mother, a dedicated researcher and a caring supervisor is inspiring and admirable. You are a role model for me and I hope to become someone like you. There were some dark moments during my PhD journey when I burst into tears in front of you. Thank you for your trust, understanding and heartfelt encouragement. You cared about my feelings, even when I tended to ignore them. Gonçalo, I will always remember your wish for me to become an elastic material, not a plastic one that breaks easily. Thank you for moulding me into such a person over the years. I appreciate all your straightforward feedback, your visionary insights and the flexibility you have allowed me to explore new ideas and directions in my research. You are like a volcano - serious and rigorous about research on the outside, but with a heart of hilarious and caring lava on the inside. Thank you for always standing by my side and supporting me whenever I needed help. I am grateful to have you as my manager and look forward to working with you for the next two years. I would also like to thank Karen for giving me the opportunity to join the Discrete Mathematics and Optimisation Group and for all the advice you gave me during our meetings.

I would like to thank all my amazing colleagues in the optimisation group: Remie, Yuki, Renfei, Jacopo, Tom, Josse, Guillermo, Lara, Esther, Naqi, Merel, Willem, Bram, Nando, Theo, Sue, Maaike, Niels, Paul and Ananth. Thank you for all the conversations, help and activities we have shared. You make my PhD journey fun

and meaningful. Now I would like to express some special thanks: Remie, thank you for your company when I was a nervous newbie in my first year. Josse, you give the firmest hugs in the world. Guillermo, thank you for remembering my birthday every year, despite the complexities of the lunar calendar. Naqi, I am grateful to have shared so many happy and bitter moments with you. Willem, I really enjoyed our summer tennis sessions. Niels, I enjoyed every one of your jokes. Paul, of all the French people I know, you are certainly the most eccentric (in a good way). I would also like to thank the support team - Dorothée, Joffrey, Xi Wei and Kees - for their unconditional help over the years.

I would like to thank all my friends. Thank you, Teng, for your unconditional support and companionship. Yves, thank you for being my mentor, my tennis coach and my friend. Thank you for always surprisingly showing up when I need a chat. Thank you, Alice, Mingxin, Yue, Xi, Yuanchen, Marco, Micky, Andres and Barbara for all the support, care and trust you have given me throughout my PhD journey. Xiaohui and Zhenheng, you are the best cooks and you made my homesickness go away. Thanks to my friends in the Transport & Planning Department - Ximing, Fei, Shuang, Zhuotong, Yiyun, Senlei, Yiru, Rina, Jie, Junhao and Jingjun - for all the help, inspiring conversations and shared experiences. Thank you, Yuxuan, for being such a great sports partner. A big thank you goes to Montaine Sanchez for the design of the cover. And to Chenming, Chang and Yu, thank you for your enduring friendship.

Finally, I would like to thank my family. To my two cats, Suisui and Niannian, thank you for bringing so much love, sweetness and joy into my life. My deepest gratitude goes to my parents - for everything.

# Contents

# Chapter 1

# Introduction

This chapter provides an introduction to this thesis. Section 1.1 presents the background information relevant to the research. Section 1.2 outlines the problem statement and the research questions. The research approach employed to address these questions is given in Section 1.3. Section 1.4 highlights the contributions of this thesis, and finally, Section 1.5 offers a brief outline of the thesis structure.

## 1.1 Background

Automated vehicles (AVs), also known as self-driving or autonomous vehicles, represent a significant shift in transportation technology. Equipped with sensors, cameras, artificial intelligence, and advanced algorithms, these vehicles can operate with minimal or no human intervention. Research has highlighted the benefits of deploying AVs, such as reducing human error and increasing road safety (Wang et al., 2020a), enhancing mobility solutions (Spieser et al., 2014; Liang et al., 2020), and reducing environmental impacts (Fagnant & Kockelman, 2014).

The Society of Automotive Engineers (SAE) has developed a widely recognised framework for vehicle automation levels, ranging from Level 0 to Level 5, with Level 5 representing full automation (On-Road Automated Driving (ORAD) committee, 2021). Level 5 AVs can perform all driving tasks, such as navigating streets, parking, and changing lanes, without human intervention. Passengers in a Level 5 AV are not required to take over driving at any time, allowing them to engage in work-related or leisure activities. This also eliminates the need for a driving licence, making this mode of transportation more accessible to everyone. These advantages of AVs are expected to drive a revolution in mobility systems.

Many researchers have investigated the potential of integrating AVs into ride-hailing

services as shared AVs (SAVs) to provide seamless door-to-door transportation in future mobility systems. SAVs, which can be centrally controlled as "moving robots", are likely to be deployed by on-demand mobility systems, benefiting service providers by eliminating driver costs and offering continuous, high-quality door-to-door service (Liang et al., 2020; Yang et al., 2020). In such a system, customers can request rides from any location using their smartphones by entering trip details like origin, destination, and departure time, and they will be matched with nearby available vehicles through the platform.

Numerous benefits have been identified for including AVs in providing ride-hailing services. According to Spieser et al. (2014), an automated mobility-on-demand (AMoD) system could meet the mobility needs of the entire population with roughly one-third of the current number of private cars. Additionally, Fagnant & Kockelman (2014) suggest that each SAV could replace approximately eleven privately owned vehicles, leading to significant reductions in energy consumption and greenhouse gas emissions. The integration of SAVs into the transportation system could also decrease parking demand, as indicated by Zhang & Guhathakurta (2017), due to higher vehicle utilisation rates and reduced reliance on private cars. This thesis focuses on this new ride-hailing system with SAVs providing on-demand mobility services and describes a comprehensive study on planning and operational decision-making from the SAV service provider's perspective.

Compared to current ride-hailing services like Uber, Lyft, and Didi, the planning and operations of future SAV services encounter numerous additional challenges. A primary challenge is the significant transformation that AVs are expected to bring to existing infrastructure. Traditional transportation networks may evolve into being part of the realm of intelligent transportation systems (ITS). Research suggests that city planners and government agencies are likely to designate specific traffic lanes (Chen et al., 2016; Liu & Song, 2019; Conceição et al., 2021) or dedicated zones (Chen et al., 2017; Madadi et al., 2020) exclusively to AVs. These AVs-only lanes/streets aim to facilitate the seamless integration of AVs into the current transportation framework, ensuring efficient and safe operations. Meanwhile, AVs-only zones are intended to maximise the benefits of AV technology by creating environments optimised for AV navigation. Initially, such zones might be established in areas prone to frequent traffic congestion—like city centres, train stations, and university campuses—to alleviate congestion through optimised traffic flow and enhanced link capacity. They might also be set up in high pedestrian traffic areas, such as commercial districts, where the precise and predictable behaviour of AVs, also in terms of low speed, can significantly lower the risk of accidents. These designated AV-only areas are expected to gradually expand, eventually transforming the entire road network into an automated and con-

nected system. In this thesis, we consider a scenario involving Level 5 AVs that can navigate freely to any destination, alongside human-driven vehicles (HVs) in mixed traffic outside of an AV-only zone. However, the AV-only zone is exclusively dedicated to AVs and prohibits HVs from entering. For ride-hailing service providers, this necessitates adaptively updating decisions such as fleet type and size with the gradual expansion of network infrastructure.

Another significant challenge in the transition to fully ITS is the coexistence of HVs with AVs on part of the road networks. Despite the significant advantages that AV technology is going to bring to mobility systems and transportation networks, the transition from traditional transportation systems to fully intelligent ones will be gradual. While AVs and HVs will most likely coexist on the same roads, as already observed in the US and China, their routing behaviours differ significantly. SAVs, centrally controlled by operators, will cooperate and follow platform guidance to benefit the overall system profit. In contrast, HVs will behave selfishly to minimise their individual costs following the well-known so-called user equilibrium principle. This mixed driving situation can be problematic, significantly reducing traffic efficiency (Yang et al., 2016). This presents a challenge for ride-hailing services: understanding different driving behaviours and studying the interactions among different traffic participants to make informed planning and operational decisions. This thesis studies the evolution of the mobility system, from a mixed driving environment to a fully automated one, to help ride-hailing services make the most profitable planning and operational decisions.

In addition to the interactions among different traffic participants, the interactions between infrastructure and vehicle routing behaviours also significantly influence the management and operational decisions of ride-hailing services. A crucial factor is the driving restrictions imposed on HVs by AVs-only zones. During the period of mixed driving, these restrictions can profoundly affect the routing behaviours of both AVs and HVs. Furthermore, link capacity restrictions and congestion effects, caused by routing large numbers of vehicles, play a vital role. AVs-only zones are specifically designed to optimally control AV flow, thereby enhancing link capacity. Consequently, when analysing future mobility systems, it is essential to consider congestion effects in the models that are used for planning and operating such systems. Although drivers/operators typically select the shortest paths for routing, capacity restrictions can make these paths suboptimal when many vehicles drive on the same route. This leads to congestion effects that trigger dynamic route choices, which in turn influence the management and operational decisions of ride-hailing services.

The emergence of this new mobility system and the resulting revolution of transportation infrastructure are expected to significantly impact passengers' travel behaviour and mode choice. The driving environment will shift from mixed traffic to a fully auto-

mated environment. However, active modes of transport, such as walking and cycling, will remain an important part of urban mobility. Even in fully automated environments with widespread adoption of SAVs, active modes are expected to remain in cities.

With this evolution, the future demand for ride-hailing services will likely differ significantly from current patterns, potentially leading to increased traffic. Therefore, it is essential for ride-hailing service providers to understand how their decisions, along with factors like congestion and diverse traveller preferences, influence mode choice behaviour. This understanding is critical for making informed decisions about fleet sizing and management, adding complexity to SAV operators' decision-making processes. This thesis aims to methodically address these challenges one by one.

## 1.2 Problem statement and research questions

The problem addressed in this thesis is a high-level planning problem from the perspective of a ride-hailing service provider. To achieve the most profitable planning decisions, it is essential to characterise the performance of a ride-hailing system during a typical day of operation. With this aim, we model the service provider's decisions at both the planning and operational levels. At the planning level, key decisions include determining the pricing strategy, fleet type and size, initial fleet distribution, and service level. At the operational level, the centralised operator matches the travel requests to available vehicles and provides optimal routing guidance to ensure timely and efficient transport. When there is no request, decisions regarding the relocation and parking of AVs need to be made. Given that a city's demand structure and transportation infrastructure evolve over time—with some areas becoming accessible only to AVs—all decisions must be adaptable to current conditions to ensure optimal profitability.

#### Research questions addressed in this thesis

*Research Question 1: How should ride-hailing service providers optimally size and manage mixed fleets of SAVs and conventional vehicles/taxis in response to the gradual expansion of AV-only zones in urban areas, considering their impact on traffic congestion?*

This question is explored in Chapters 2 and 3.

*Research question 2: How can we model the interactions between different routing behaviours—specifically, privately-owned HVs following the user equilibrium (UE) and centrally dispatched vehicles/taxis following the system optimum (SO)? How do these interactions influence the optimal sizing and management of the fleets?*

This question is addressed in Chapter 3.

*Research question 3: How can existing models be adapted to incorporate endogenous demand to plan and operate an SAV service?*

This question is answered in Chapter 4.

*Research question 4: What are the optimal pricing strategies for SAV services, considering the interplay between demand and supply variables, congestion effects, and the heterogeneous income levels of travellers?*

This question is answered in Chapter 5.

## 1.3 Research approach

Mathematical optimisation and simulation are two primary approaches used by researchers to study the optimal management and operations of ride-hailing services (Liang et al., 2020; Pinto et al., 2020; Wei et al., 2022). Simulation approaches can replicate complex scenarios by accounting for the diverse behaviours of road users and monitoring their dynamic interactions. However, these techniques are typically time-consuming, as they require running a large number of simulations to evaluate system performance by exploring combinations of many decisions to find optimal solutions. Therefore, simulation is more suited to evaluate multiple scenarios rather than pinpointing the optimal combination of values assigned to decision variables in a combinatorically complex system.

Mathematical optimisation seeks the best solution from a set of available alternatives, aiming to minimise or maximise an objective function. This function may represent components that need to be minimised—such as cost, time, or distance—and/or maximised—such as profit or efficiency. The variables in these models are the controllable elements used to achieve these goals and often come with constraints that define the feasible region for viable solutions.

The questions we aim to address in this thesis are more suitable to be tackled using mathematical optimisation approaches. These include classical optimisation methods—such as linear programming, nonlinear programming, integer programming, dynamic programming, and stochastic programming—as well as alternative approaches like heuristics and metaheuristics. Classical optimisation techniques aim to find optimal solutions which can be challenging for large-scale and complex problems. Heuristics and metaheuristics offer more flexibility in exploring diverse solution spaces and adapting to various problem structures, often providing satisfactory solutions within a shorter timeframe. Unlike classical methods, heuristics and metaheuristics do not guarantee optimality or convergence under strict conditions, but they are well-suited

for practical problem-solving where a feasible solution is needed in a reasonable/short time. In this thesis, we explore various solution methods, integrating them to develop tailored approaches for the studied problems. The proposed methods leverage the strengths of different techniques to enhance overall effectiveness.

Table 1.1 gives an overview of the elements modelled in each chapter of this thesis. In Chapter 2, we develop a flow-based vehicle routing model to determine the optimal fleet size of SAVs and conventional taxis. This model considers the gradually increasing coverage of dedicated AV-only zones. Traffic congestion is incorporated through flow-dependent travel times. We test two service regimes: the User Preference Mode (UPM), where passengers select their vehicle type based on personal preferences, and the System Profit Mode (SPM), where the taxi company assigns vehicles to maximise profits. The model is formulated as a mixed integer linear programming model and solved using a state-of-the-art solver.

*Table 1.1: Overview of research elements and decisions*

| | Chapter 2 | Chapter 3 | Chapter 4 | Chapter 5 |
|---|---|---|---|---|
| Driving environment | AVs and HVs mixed driving | AVs and HVs mixed driving | Fully automated driving | Fully automated driving |
| Evolving AVs-only zone? | Yes | Yes | No | No |
| Traffic congestion modelling? | Yes | Yes | Yes | Yes |
| Mixed routing behaviour modelling? | No | Yes | No | No |
| Mode choice modelling? | No | No | Yes | Yes |
| Main decisions | Fleet type and size | Fleet type and size | Fleet size and initial distribution, service quality | Pricing strategies, fleet size and initial distribution, service quality |

In Chapter 3, we introduce a bi-level framework that captures the mixed routing behaviour of vehicles (both HVs and AVs) and endogenous traffic congestion at the lower level, while the upper level determines the fleet size to maximise profit. This framework is tackled using a parallel genetic algorithm, embedded with a tailored iterative algorithm for solving the lower-level model.

In Chapter 4, we formulate a mixed-integer non-linear programming model that addresses congestion effects and the mode choices of urban travellers across different income classes, between SAVs and bicycles. Travellers' preferences for both transport

modes are modelled using a binary logit model, and congestion effects are described through dynamically varying travel times based on traffic flow in a non-linear manner. Additionally, we explore two types of accept/reject mechanisms for the service operator (mandatory versus non-mandatory acceptance), which influence an endogenously determined acceptance rate affecting travellers' willingness to use SAV services. The computational challenges posed by the non-linear and non-convex nature of the model are addressed through a reformulation and the use of outer-inner approximation methods combined with a breakpoint generation algorithm. Then, the reformulated model is solved using a state-of-the-art solver.

In Chapter 5, we extend the model introduced in Chapter 4 to determine the optimal pricing and fleet management decisions under three distinct pricing strategies: base fare plus distance-based fare, distance-based fare only, and income class-based fare. We then develop three unique solution algorithms to address the model's complex non-linearities from different perspectives. These approaches include linearisation techniques, hybrid metaheuristic-based optimisation, and hybrid Bayesian optimisation-based methods. A comparative analysis of these methods is conducted.

## 1.4 Thesis contributions

This thesis makes contributions from both scientific and practical perspectives.

From a scientific perspective, it enhances the well-explored area of fleet sizing and management for on-demand mobility services by incorporating several new elements. These elements include the evolution of infrastructure, particularly the emergence and expansion of AV-only zones; the interactions of different routing behaviours, such as centrally dispatched taxis following the SO and privately-owned HVs following the UE; the endogenous congestion arising from the routing of both AVs and HVs; the consideration of endogenous demand; operators' accept/reject decisions; and the development of pricing strategies that consider the heterogeneous preferences of travellers. Additionally, this thesis introduces novel mathematical models designed to optimise planning and operational decisions to maximise the profitability of ride-hailing services. Tailored solution algorithms have also been developed and compared to effectively solve these complex mathematical models.

On the practical side, this thesis offers methodologies for ride-hailing service providers to make the most profitable decisions at both the planning and operational levels. It also demonstrates the benefits of utilising AVs for ride-hailing services. Moreover, it provides valuable managerial insights for ride-hailing service providers, city planners, and government officials regarding the potential impact of AV-related infrastructure.

The thesis further explores the complex interactions between traffic participants and infrastructure, as well as the interplay between supply and demand, thereby deepening the understanding of the dynamics within SAV systems.

## 1.5 Thesis outline

Figure 1.1 provides an overview of the thesis structure, which consists of six chapters. Chapter 1 introduces the background, problem statement, research questions, research approaches, and both scientific and practical contributions. Chapters 2 to 5 include the main content of the published and under-review papers.

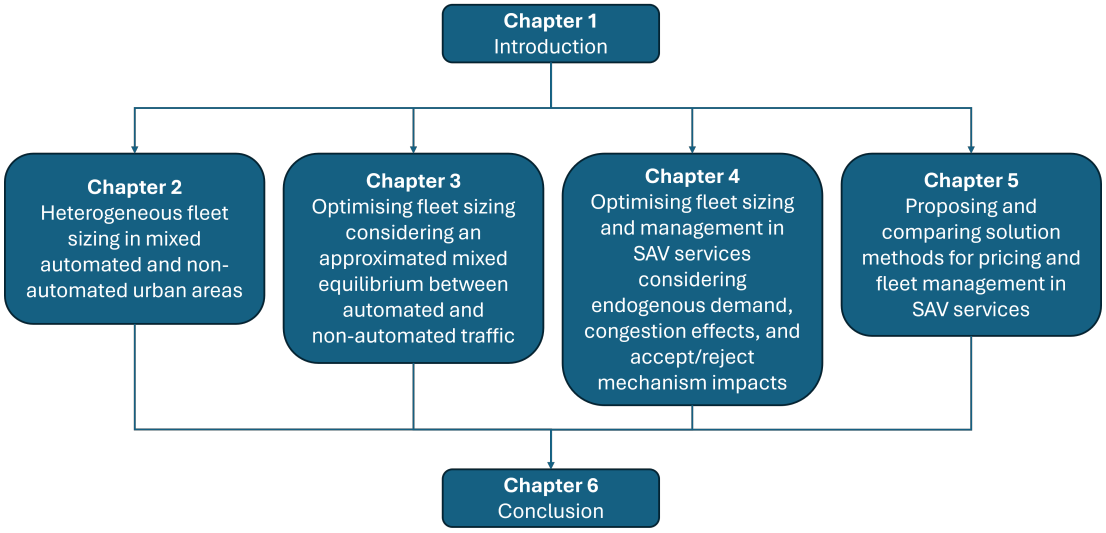



*Figure 1.1: Overview of thesis structure*

Chapter 2 addresses the heterogeneous fleet sizing problem in the context of the emergence of AVs-only zones. The model formulations are applied to a case study using a large, simulated network, providing insights into the performance of a heterogeneous taxi system on a hybrid network. This chapter has been published in Transportation Research Procedia:

> Fan, Q., Van Essen, J. T., & Correia, G. H. (2022). Heterogeneous fleet sizing for on-demand transport in mixed automated and non-automated urban areas. *Transportation Research Procedia*, 62, 163-170.

Chapter 3 introduces a bi-level framework to capture the mixed routing behaviour of vehicles and endogenous traffic congestion when making fleet sizing and management decisions. The proposed solution methods are tested using instances based on a small network and the network of the city of Delft, the Netherlands, to investigate the impacts of AVs-only zones on traffic and ride-hailing operations. This chapter has been published in the European Journal of Operational Research:

> Fan, Q., van Essen, J. T., & Correia, G. H. (2024). A bi-level framework for heterogeneous fleet sizing of ride-hailing services considering an approximated mixed equilibrium between automated and non-automated traffic. *European Journal of Operational Research*, 315(3), 879-898.

Chapter 4 envisions a fully automated driving environment where AVs replace private cars and offer public on-demand mobility services to meet the mobility needs of city residents. Our proposed method is applied to the case study of the city of Delft in the Netherlands. Additionally, we conduct scaling analyses on three simulated networks of varying sizes and demand profiles to demonstrate the effectiveness of our proposed method. This chapter has been published in Transportation Research Part C:

> Fan, Q., van Essen, J. T., & Correia, G. H. (2023). Optimising fleet sizing and management of shared automated vehicle (SAV) services: A mixed-integer programming approach integrating endogenous demand, congestion effects, and accept/reject mechanism impacts. *Transportation Research Part C: Emerging Technologies*, 157, 104398.

Chapter 5 explores optimal pricing strategies for the system studied in Chapter 4 and develops three different solution algorithms to address the non-linearity from various perspectives. The performance of these algorithms is compared to assess their efficacy. This chapter has been submitted for publication.

> Fan, Q., van Essen, J. T., & Correia, G. H. (Under review). Solution methods for pricing and fleet management in shared automated vehicle services considering supply-demand dynamics, congestion, and income heterogeneity.

Chapter 6 discusses the main conclusions and directions for future research.

# Chapter 2

# Heterogeneous fleet sizing for on-demand transport in mixed automated and non-automated urban areas

---

The era of intelligent transportation with automated vehicles (AVs) is coming. Nonetheless, the transition to this system will be a gradual process. On the one hand, some zones in the city may be dedicated to AVs with a fully intelligent traffic management system geared toward high performance. On the other hand, automated and conventional vehicles may have to be allowed to drive in the remaining zones of the urban network in a transition stage. In this chapter, we consider a situation where AVs are deployed by a taxi operating company to serve door-to-door travel requests. Facing this transition period, a strategic flow-based vehicle routing model is developed to determine the optimal fleet size of automated and conventional taxis as a function of the gradually increasing coverage of the AVs-only dedicated area. The developed model formulations are applied to a case study of a large toy network.

This chapter is structured as follows: Section 2.1 gives the introduction to the studied problem. Section 2.2 establishes a flow-based routing model for heterogeneous vehicles in a mixed automated and non-automated zone network incorporating different service regimes. Section 2.3 presents the numerical results of a case study with a large toy network. Section 2.4 draws our conclusions.

---

## 2.1 Introduction

In recent years, various technologies for automated driving have been developed and extensively tested, which leads to believe that automated vehicles (AVs) are coming to the market soon. This will revolutionise people's travel patterns. For example, the emergence of shared automated vehicles (SAVs) will challenge the usage of privately-owned cars and public transport as they can provide on-demand door-to-door service to meet personal mobility needs. Novel business models using SAVs may emerge and be deployed globally, providing app-based on-demand service. This will allow passengers to make online requests providing their desired trip information including the origin and destination. The operating system will then assign passengers' requests to vehicles and determine the vehicles' route in the transportation network.

Many papers focus on optimising the profit from the operator's perspective (Liang et al., 2017); Liang et al. (2020). However, most of them consider the situation where all travel requests are served by SAVs. By this, they ignore the fact that the transition to such an intelligent automated transportation system will be a slow and gradual process. Many researchers hold the view that some critical locations or zones, such as the city centre or locations where it is easy to have traffic bottlenecks, are likely to be the first to be dedicated to AVs thus establishing AVs-only zones to improve traffic efficiency (Chen et al., 2017)). Within an AVs-only zone, AVs will follow the route guidance given by the operating company realising a fully automated driving environment. The human-driven vehicles would therefore be prohibited from entering the AVs-only zone to avoid randomness brought by human drivers. In the remaining part of the network, AVs and conventional vehicles (CVs) used as shared taxis are very likely to cooperate to satisfy the travel requests. Considering the gradual expansion process of intelligent infrastructure in the city, it is much more realistic at this point to build models that consider vehicle routing in mixed traffic conditions (automated and human-driven).

To determine a new strategy for a taxi company when facing the upcoming SAVs era, we develop a strategic flow-based vehicle routing model in a time-space network to determine the optimal fleet size of automated taxis (ATs) and make adjustments to the existing fleet size of conventional taxis (CTs) as the coverage of the AVs-only zone expands. We assume that ATs at level 5 automation and CTs co-exist to serve the total mobility demand in a city. Due to the restriction of an existing AVs-only zone, CTs cannot drive in the whole network whilst ATs can drive in both the AVs-only zone and outside the AVs-only zone. Traffic congestion is considered in the model by making travel times dependent on the vehicle flows. Furthermore, two service regimes are considered. The first one is a User Preference Mode (UPM) where passengers

are allowed to choose the vehicle type (AT or CT) if their origin and destination are both outside the AVs-only zone. The user's preference towards the vehicle type makes sense since many people may prefer low levels of automation as they worry about the potential risk of the AVs without human supervision (Ha et al., 2020). The second one is a System Profit Mode (SPM) in which the taxi company will take charge of the vehicle assignment to maximise the profit.

## 2.2 Methodology: mixed-integer linear programming (MILP) models

The aim of this chapter is to develop a method to determine the fleet size of ATs and CTs for a taxi company in a city while the AVs-only zone is expanding. To characterise the performance of a taxi system during a general day of operations, one must be able to assign travellers to taxis and to determine vehicle routes under a mixed driving environment subject to AVs-only zone constraints. The interplay between the route choice of vehicles and dynamic travel time considering traffic congestion is also incorporated in this model.

We have the following assumptions for this future scenario: (1) ATs are allowed to move empty in the network without a human driver; (2) Vehicles can only park at certain nodes which are defined as the parking depots provided by the taxi company; (3) To keep a high-quality service, the taxi company cannot reject any travel request; (4) Only the taxi company uses AVs, so travellers are assumed not to be able to use private AVs; (5) The background traffic flow generated by privately-owned human-driven cars outside the AVs-only zone is simplified. Constants are used to represent the average value of such background traffic. Considering the driving restrictions imposed on CVs, trips received by the on-demand mobility system should be assigned to the appropriate type of vehicle based on three types of trips: 1) a trip with origin and destination inside the AVs-only zone which should be served by an AT; 2) a trip with origin and destination outside of the AVs-only zone which can be served by either an AT or CT; 3) a trip with origin/destination inside the AVs-only zone and destination/origin outside the AVs-only which should be served by an AT. Next, the mathematical models for UPM and SPM are presented.

*Table 2.1: Notation.*

| Notation | Description |
|---|---|
| **Sets** | |
| $T$ | Set of time instants in the operation period. |
| $N$ | Set of nodes. |
| $L$ | Set of road links between nodes in set $N$. |
| $G$ | Set of links in the time-space network. |
| $M$ | Set of vehicle types, with option 1 being the conventional taxi $(CT)$ and option 2 being the automated taxi $(AT)$. |
| $R$ | Set of groups of requests, where each group of requests $r \in R$ has the same origin, destination, desired departure time, and latest arrival time at the destination. |
| $N^m$ | Set of nodes that can be used by vehicles of type $m \in M$ with $N^m \subseteq N$. CTs can use the nodes outside the AVs-only zone and the nodes located at the border of the AVs-only zone; ATs can use all the nodes. |
| $N_P^m$ | Set of nodes allowing parking for vehicles of type $m \in M$ with $N_P^m \subseteq N^m$. |
| $G^m$ | Set of links that can be used by vehicles of type $m \in M$ in the time-space network. |
| $R^m$ | Set of groups of requests served by vehicles of type $m \in M$ with $R^m \subseteq R$. |
| **Parameters** | |
| $n^r$ | Total number of requests for group of requests $r \in R$. |
| $o^r$ | Origin node for group of requests $r \in R$. |
| $d^r$ | Destination node for group of requests $r \in R$. |
| $a^r$ | Desired departure time for group of requests $r \in R$. |
| $b^r$ | Latest arrival time for group of requests $r \in R$. |
| $f_{ijt}$ | Background traffic flow on road link $(i, j) \in L$ at time instant $t \in T$. |
| $Q_{ij}$ | Capacity of road link $(i, j) \in L$ in vehicles per time unit. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of road link $(i, j) \in L$ in vehicles that fit on the road link from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |
| $t_{ij}^{\max}$ | Maximum travel time on road link $(i, j) \in L$. |
| $t_{ij}^{\min}$ | Minimum travel time on road link $(i, j) \in L$. |
| $p^0$ | Initial base fare in euros for using the taxis. |
| $p^m$ | Price per kilometre in euros/km for using vehicle type $m \in M$. |
| $co^m$ | Unit driving operational cost in euros/km for vehicle type $m \in M$. |
| $cd$ | Delay cost in euros per time instant. |
| $cp$ | Salary of a driver in euros per time instant. |
| $cf^m$ | Depreciation cost in euros per vehicle per time step for using vehicle type $m \in M$. |
| $l_{ij}$ | Length of road link $(i, j) \in L$. |
| $std^r$ | Shortest travel distance for group of requests $r \in R$. |
| $stt^r$ | Shortest travel time assuming free flow speed for group of requests $r \in R$. |
| $s$ | Total number of time instants in the operation period. |

| | |
|---|---|
| **Decision variables** | |
| $PF^{r}_{i_{t_1} j_{t_2}}$ | Passenger flow in group of requests $r \in R^m$ served by vehicle type $m \in M$ in road link $(i, j)$, from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G^m$, $a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| $PF^{rm}_{i_{t_1} j_{t_2}}$ | Passenger flow in group of requests $r \in R$ served by vehicle type $m \in M$ in road link $(i, j)$ from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G^m$, $a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| **Auxiliary variables** | |
| $V^m$ | Taxi fleet size of type $m \in M$. |
| $E^{rt}$ | Total number of passengers in group of requests $r \in R^m$ for vehicle type $m \in M$ arriving at time $t \in T$. |
| $TF^{m}_{i_{t_1} j_{t_2}}$ | Total number of taxis of type $m \in M$ in road link $(i, j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G^m$. |
| $TP^{m}_{i_t}$ | Total number of taxis of type $m \in M$ parking at node $i \in N^m_P$ from time instant $t$ to $t+1$, with $t \in T$. |
| $X_{i_{t_1} j_{t_2}}$ | Binary variable which is 1 when vehicles travel in road link $(i, j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. |
| $F_{ijt}$ | Traffic flow on road link $(i, j) \in L$ at time instant $t \in T$. |
| $P^{rm}$ | Total number of requests in group of requests $r \in R$ served by vehicle type $m \in M$. |
| $E^{rmt}$ | Total number of passengers in group of requests $r \in R$ served by vehicle type $m \in M$ arriving at time $t \in T$. |

### 2.2.1 User preference mode (UPM)

As explained before, in this on-demand mobility service system, we consider a service regime called 'user preference mode' in which travellers can choose the vehicle type by themselves if the requests can be served by both types of vehicles. While for the trips of type 1 and type 3, there is no option for the passengers and they always have to use an AT. In this case, we assume that all the passengers have known the available options and will adjust their behaviour to the on-demand mobility system. Thus, the requests for the vehicle type will always be feasible and no request will be rejected. This mode takes users' preferences into account which will increase the users' satisfaction with the on-demand mobility service system. A mixed-integer linear programming (MILP) model is developed with the objective of maximising the total profit of the whole system.

#### Base formulation

The travel requests in group $r \in R^m$ have the same origin and destination node. Even though the vehicles serving group $r \in R^m$ depart at the same time, they may arrive at

the destination at different times. Constraints (2.1)-(2.3) make sure that, for each group $r \in R^m$, the passenger flows departing from the origin node $o^r$ at time $a^r$ and arriving at the destination node $d^r$ should be equal to the total number of requests.

$$\sum_{j_t | \left(o^r_{a^r}, j_t\right) \in G^m} PF^r_{o^r_{a^r} j_t} = n^r, \forall r \in R^m, m \in M \tag{2.1}$$

$$\sum_{t \in T | a^r + stt^r \leq t \leq b^r} E^{rt} = n^r, \forall r \in R^m, m \in M \tag{2.2}$$

$$E^{rt} = \sum_{i_{t_1} | \left(i_{t_1}, d^r_t\right) \in G^m} PF^r_{i_{t_1} d^r_t}, \forall r \in R^m, m \in M, t \in T \tag{2.3}$$

Constraints (2.4) and (2.5) ensure that the passenger flows are generated in the origin node and absorbed in the destination node. The passenger flow conservation in the intermediate nodes is described in Constraints (2.6). The total number of passengers on road link $(i, j)$ travelling from time instant $t_1$ to time instant $t_2$, should be less than or equal to the total number of taxis on the same link as some taxis might drive without passengers, as in Constraints (2.7).

$$\sum_{j_{t_2} | \left(d^r_{t_1}, j_{t_2}\right) \in G^m} PF^r_{d^r_{t_1} j_{t_2}} = 0, \forall r \in R^m, m \in M, t_1 \in T, a^r + stt^r \leq t_1 \leq b^r \tag{2.4}$$

$$\sum_{i_{t_1} | \left(i_{t_1}, o^r_{t_2}\right) \in G^m} PF^r_{i_{t_1} o^r_{t_2}} = 0, \forall r \in R^m, m \in M, t_2 \in T, a^r \leq t_2 \leq b^r \tag{2.5}$$

$$\sum_{j_{t_0} | \left(j_{t_0}, i_{t_1}\right) \in G^m} PF^r_{j_{t_0} i_{t_1}} = \sum_{j_{t_2} | \left(i_{t_1}, j_{t_2}\right) \in G^m} PF^r_{i_{t_1} j_{t_2}}, \forall r \in R^m, m \in M, t_1 \in T, t_0 < t_1 < t_2, i \in N^m, i \neq o^r, i \neq d^r \tag{2.6}$$

$$\sum_{r \in R^m} PF^r_{i_{t_1} j_{t_2}} \leq TF^m_{i_{t_1} j_{t_2}}, \forall \left(i_{t_1}, j_{t_2}\right) \in G^m, m \in M \tag{2.7}$$

At the beginning of the service period, the total number of taxis driving on road link $(i, j)$ plus the total number of taxis parked at depot $i \in N^m_P$ should be equal to the fleet size of AT and CT as specified in Constraints (2.8). Constraints (2.9) and (2.10) describe the vehicle flow equilibrium for the nodes that allow and do not allow vehicle parking, respectively. Outside the AVs-only zone, the traffic flow of road link $(i, j)$ at time instant $t$ is calculated by the background traffic flow generated by privately owned CVs together with the flow generated by ATs and CTs, while in the AVs-only zone, the

value of background traffic flow is zero and the traffic flow will be generated only by ATs. Constraints (2.11) calculate the traffic flow on every link.

$$\sum_{(i_0,j_t)\in G^m} TF^m_{i_0 j_t} + \sum_{i\in N^m_P} TP^m_{i_0} = V^m, \forall m \in M \tag{2.8}$$

$$\sum_{(j_{t_1},i_t)\in G^m | t_1<t} TF^m_{j_{t_1} i_t} + TP^m_{i_{t-1}} = \sum_{(i_t,j_{t_2})\in G^m | t<t_2} TF^m_{i_t j_{t_2}} + TP^m_{i_t}, \forall t \in T, i \in N^m_P, m \in M \tag{2.9}$$

$$\sum_{(j_{t_1},i_t)\in G^m | t_1<t} TF^m_{j_{t_1} i_t} = \sum_{(i_t,j_{t_2})\in G^m | t<t_2} TF^m_{i_t j_{t_2}}, \forall t \in T, i \in N^m \setminus N^m_P, m \in M \tag{2.10}$$

$$F_{ijt} = f_{ijt} + \sum_{m\in M} \sum_{t_2\in T | (i_t,j_{t_2})\in G^m} TF^m_{i_t j_{t_2}}, \forall t \in T, (i,j) \in L \tag{2.11}$$

**Traffic congestion**

We use the formulation of Van Essen & Correia (2019) to include traffic congestion in our model. Based on the Bureau of Public Roads (BPR) function, we calculate the spatial capacity $C_{i_{t_1} j_{t_2}}$ of road link $(i,j)$ from time instant $t_1$ to time instant $t_2$ by Equation (2.12). Note that if the difference between $t_2$ and $t_1$ equals the minimum travel time of road link $(i,j)$, the value of $C_{i_{t_1} j_{t_2}}$ will be zero. To avoid this, we add in this case 0.5 to $t_2$ to obtain a nonzero value. To match the road link flow with the spatial link capacity, binary variables $X_{i_{t_1} j_{t_2}}$ are introduced. Constraints (2.13) describe the allowed total flow on road link $(i,j)$ at time instant $t_1$. Constraints (2.14) ensure that at most one travel time for road link $(i,j)$ starting from time instant $t_1$ can be chosen. Constraints (2.15) ensure that there is only vehicle flow from time instant $t_1$ to at most one time instant $t_2$. The first-in first-out Constraints (2.16) ensure that the vehicle that enters the road link first will leave the road link first, which means that vehicles cannot pass one another when driving on a link.

$$C_{i_{t_1} j_{t_2}} = (t_2 - t_1) Q_{ij} \left( \frac{1}{a} \left( \frac{t_2 - t_1}{t^{\min}_{ij}} - 1 \right) \right)^{\frac{1}{b}}, \forall (i_{t_1}, j_{t_2}) \in G \tag{2.12}$$

$$\sum_{t_2\in T} \left\lceil C_{i_{t_1} j_{(t_2-1)}} \right\rceil X_{i_{t_1} j_{t_2}} \leq F_{ijt_1} \leq \sum_{t_2\in T} \left\lfloor C_{i_{t_1} j_{t_2}} \right\rfloor X_{i_{t_1} j_{t_2}}, \forall (i,j) \in L, t_1 \in T \tag{2.13}$$

$$\sum_{t_2 | (i_{t_1}, j_{t_2})\in G} X_{i_{t_1} j_{t_2}} \leq 1, \forall (i,j) \in L, t_1 \in T \tag{2.14}$$

$$X_{i_{t_1} j_{t_2}} \geq \frac{\sum_{m \in M} TF^m_{i_{t_1} j_{t_2}}}{C_{i_{t_1} j_{t_2}}}, \forall (i_{t_1}, j_{t_2}) \in G \tag{2.15}$$

$$t_1 + \sum_{t \in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \sum_{t \in T} X_{i_{t_2} j_t} (t - t_2) + M \left(1 - \sum_{t \in T} X_{i_{t_2} j_t}\right), \forall t_1 < t_2 \in T, (i, j) \in L \tag{2.16}$$

**Objective function**

From the operator's point of view, the aim is to maximise the total profit of the whole system, which includes the taxi fares paid by passengers, the operational cost (including fuel, cleaning, maintenance, etc.) of the fleet, the delay penalisation, the salaries for drivers, and the depreciation cost of the taxis. The taxi fares paid by passengers are constant in Equation (2.17) as the number of requests served by ATs and CTs are known beforehand. This term is included in the objective function to be able to compare with the SPM in which the vehicle type serving a request is determined by the model.

$$\begin{aligned} \max \sum_{m \in M} \sum_{r \in R^m} \left(p^0 \cdot n^r + p^m \cdot n^r \cdot std^r\right) - \sum_{m \in M} co^m \cdot \left(\sum_{(i_{t_1}, j_{t_2}) \in G^m} TF^m_{i_{t_1} j_{t_2}} \cdot l_{ij}\right) \\ - cd \cdot \sum_{m \in M} \sum_{r \in R^m} \left(\sum_{t \in T} E^{rt} \cdot t - a^r \cdot n^r - stt^r \cdot n^r\right) - s \cdot cp \cdot V^{CT} - \sum_{m \in M} s \cdot cf^m \cdot V^m \end{aligned} \tag{2.17}$$

## 2.2.2 System profit mode (SPM)

The UPM may lead to higher customer satisfaction, but a lower revenue for the operating company, because of additional relocations of vehicles. From the operator's point of view, the best way to maximise the system profit is to decide on the vehicle to assign to each client. In this mode, the set of groups of requests $R$ can be divided into three subsets $R_1, R_2, R_3$, representing the set of groups of requests of type 1, type 2, and type 3, respectively. For trips of type 1 and 3, the vehicle type is known beforehand, whereas the mode of vehicles serving trips of type 2 is determined by the model.

The requests in group $r \in R_2$, which have the same trip information, might be assigned to different vehicle types to maximise the system profit. Constraints (2.18) ensure that the number of requests $P^{rm}$ in group $r \in R$ served by vehicle type $m \in M$ in total equals the number of requests in group $r \in R$. Constraints (2.19) impose that the

requests of type 1 and 3 should only be served by ATs. Constraints (2.1) and (2.2) are replaced by Constraints (2.20) and (2.21). For each group of requests $r \in R$ served by different vehicle types $m \in M$, the variables $E^{rt}$ and $PF^{r}_{i_{t_1} j_{t_2}}$ in Constraints (2.3)-(2.7) should be replaced by $E^{rmt}$ and $PF^{rm}_{i_{t_1} j_{t_2}}$, respectively. The objective function for the SPM should also be modified. When calculating the total taxi fares paid by passengers and the delay penalisation, the total number of requests $n^r$ in group $r \in R$ in Equation (2.17) should be replaced by $P^{rm}$. The set of groups of requests $R^m$ should be replaced by $R$.

$$\sum_{m \in M} P^{rm} = n^r, \forall r \in R \tag{2.18}$$

$$P^{rm} = 0, \forall r \in R_1 \cup R_3, m = CT \tag{2.19}$$

$$P^{rm} = \sum_{j_t | \left(o^r_{a^r}, j_t\right) \in G^m} PF^{rm}_{o^r_{a^r} j_t}, \forall r \in R, m \in M \tag{2.20}$$

$$P^{rm} = \sum_{t \in T | a^r + stt^r \leq t \leq b^r} E^{rmt}, \forall r \in R, m \in M \tag{2.21}$$

## 2.3 Computational results

We test the models on a large toy network consisting of 64 nodes and 112 links (two-way circulation allowed). Each road link has an equal length of 2 kilometres. We use several networks with different AVs-only zone coverage rate of the road network, namely 10%, 30%, 50%, 70%, and 90%, as shown in Fig. 1. The time step is set to 2.5 minutes. The shortest travel time and the longest travel time for each road link are 2.5 minutes and 10 minutes, respectively. The background traffic flow outside the AVs-only zone is generated randomly within the capacity restriction. 600 requests and their trip information including origins, destinations, departure time, latest arrival time, number of passengers in each group, shortest distance, and shortest travel time, are generated randomly with equal probability, to emulate the travel demands in the peak hour. Assuming that the acceptance of users towards AVs in level 5 is low, more than 80% of the requests with a preference for CVs are generated. In this case study, the value of the parameters are as follows: the initial base fare $p^0$ for using the taxis is 2.66 euros and is based on the price rate of a ride-hailing company in the Netherlands; the prices $p^m$ for using CTs and ATs are 1.95 euros/km and 1.8 euros/km, respectively; the unit operational costs $co^m$ for using CTs and ATs are 0.24 euros/km and 0.32 euros/km, respectively, calculated according to the methodology proposed by Bösch et al. (2018); the delay cost $cd$ is 0.5 euros/time instant based on Liang et al. (2020); the salary

$cp$ of a driver is 10 euros/hour according to the minimum wage in the Netherlands, resulting in 0.42 euros/time instant; the depreciation costs $cf^m$ for CTs and ATs are 0.04 and 0.05 euros/time instant/vehicle, respectively, which is calculated as the price of a taxi divided by its statutory lifespan. The estimation parameters $a$ and $b$ of the BPR function are set to 2 and 4, respectively, based on Van Essen & Correia (2019). We solve these models using the Python interface of Gurobi 9.0.2 on an Intel(R) Core(TM) i5-6500 CPU @3.6GHz 8.0GB RAM computer. A comparison of the UPM and SPM in scenarios with a different coverage rate of the AVs-only zone is given in Table 2.2.

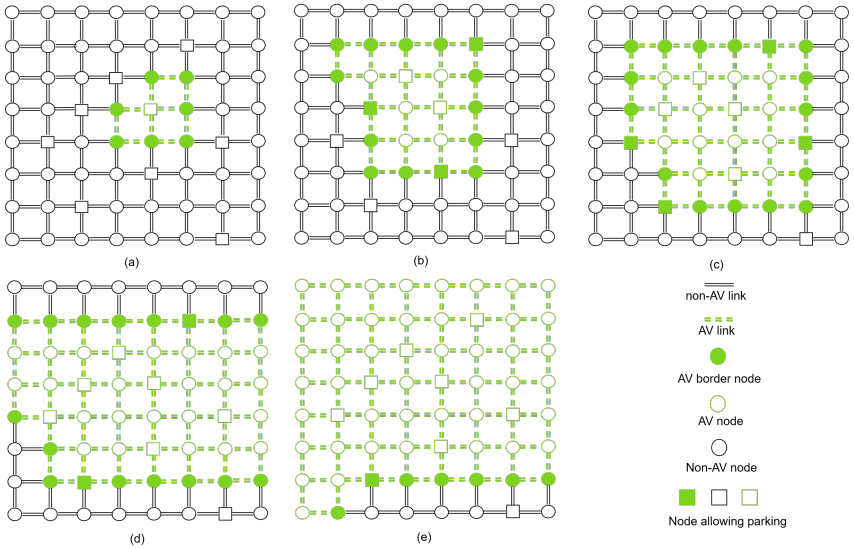



*Figure 2.1: Networks with different AVs-only zone size: (a) 10% (b) 30% (c) 50% (d) 70% (e) 90%*

The total profit increases with the coverage rate of the AVs-only zone in most cases. This happens, first of all, because the privately-owned CVs are not allowed to drive inside the AVs-only zone, resulting in a lower delay cost caused by traffic congestion when the AVs-only zone enlarges. Secondly, in the SPM, the total travel distance decreases with the increase of the AVs-only zone coverage, as less detour and relocation kilometres are needed when the congestion effect diminishes. For UPM, the total travel distance increases at the early stage as the CTs have to detour more on the road network because of the driving restriction. When the AVs-only zone is big enough, fewer CTs are required and their total travel distance decreases correspondingly. In addition to this, a smaller fleet size of taxis is needed when the AVs-only zone expands, as fewer CTs are needed and the usage rate of ATs increases. This leads to less depreciation

*Table 2.2: Optimal results for the different coverage rates of the AVs-only zone.*

| AVs-only zone coverage rate % | Service mode | Obj. value | Total fleet size | AT fleet size | CT fleet size | Satisfied requests by ATs | Satisfied requests by CTs | Total travel distance (km) | Relocation distance (km) | Detour distance (km) | Delayed time (time step) | CPU time (minutes) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10% | UPM | 4526.6 | 160 | 20 | 140 | 40 | 560 | 4870 | 1354 | 276 | 160 | 6.9 |
| | SPM | 5836.0 | 123 | 123 | 0 | 600 | 0 | 4348 | 1072 | 36 | 18 | 12.5 |
| 30% | UPM | 3831.5 | 260 | 100 | 160 | 200 | 400 | 4980 | 1444 | 296 | 229 | 5.8 |
| | SPM | 5846.7 | 122 | 122 | 0 | 600 | 0 | 4332 | 1068 | 24 | 12 | 10.6 |
| 50% | UPM | 3917.4 | 260 | 120 | 140 | 320 | 280 | 5122 | 1746 | 136 | 188 | 5.5 |
| | SPM | 5852.5 | 122 | 122 | 0 | 600 | 0 | 4320 | 1064 | 16 | 8 | 9.9 |
| 70% | UPM | 4906.3 | 181 | 121 | 60 | 480 | 120 | 4910 | 1570 | 100 | 170 | 6.1 |
| | SPM | 5856.0 | 121 | 121 | 0 | 600 | 0 | 4292 | 1044 | 8 | 4 | 8.1 |
| 90% | UPM | 5283.2 | 160 | 120 | 40 | 520 | 80 | 4544 | 1292 | 12 | 86 | 6.3 |
| | SPM | 5856.0 | 121 | 121 | 0 | 600 | 0 | 4292 | 1044 | 8 | 4 | 8.1 |

cost of the taxi fleet. A larger AVs-only zone means that more requests can be served by ATs instead of CTs which will also reduce the salary cost. The total profit does not decrease a lot in the SPM with different AVs-only zone sizes, as traffic congestion is the only factor that impacts the profit and the fleet size. When the AVs-only zone enlarges up to 70%, there is no variation in the performance of this model because the congestion effect has already been greatly reduced. Even though there exist privately-owned vehicles, ATs can always make use of the links within the AVs-only zone and find an alternative path to reduce the congestion effect. If the coverage rate of the AVs-only zone is 100%, there is no difference between UPM and SPM as all the requests will be assigned to ATs. An exception happens in the UPM when the AVs-only zone enlarges in the initial stage. The total profit falls steeply, with the coverage of the AVs-only zone increasing from 10% to 30%. This is due to the longer relocation and detour distance of CTs. Without permission to drive across the AVs-only zone, the CTs should detour more to satisfy the users' requirements and drive empty for a longer distance for the next requests. Once a CT detours, a delay cost is obtained as the CT should spend more time serving the requests. Thus, more CTs are needed to fulfil all the mobility needs, resulting in a higher salary cost for drivers. When comparing the UPM with the SPM, it is evident that the SPM can bring more profit than the UPM. All the requests are assigned to ATs in SPM, as ATs services are more profitable due to less detour and relocation cost, no salaries for drivers, and less delay penalization. In terms of the computation time, the SPM takes longer than the UPM as also the vehicle type for each request in $R_2$ needs to be determined.

## 2.4 Conclusions and future research

In this chapter, we introduced an MILP model to determine the fleet size under different service regimes and study the impact of an AVs-only zone on the taxi service system performance. In general, ATs can bring more profit than CTs. The operating company should deploy more ATs when the AVs-only zone emerges if they do not consider the users' preference towards the type of taxis. UPM brings less profit than SPM but can satisfy passengers' demand if they prefer to ride in a CV. In the long run, it is still worthwhile to consider users' preference. At the early stage, the emergence of the AVs-only zone will lead to a longer detour and relocation distance for CTs. So a well-designed construction strategy of the AVs-only zone can be beneficial to help diminish the negative effects for conventional human-driven vehicles. When the coverage rate of the AVs-only zone is relatively large, the traffic congestion will be largely reduced and the taxi operating company can gain more profit by using ATs. Further research should be done in real case studies, considering the impact of the AVs-only zone on privately-owned vehicles in a global view. Also, a sensitivity analysis of the parameters should be performed.

# Chapter 3

# A bi-level framework for heterogeneous fleet sizing considering an approximated mixed equilibrium between automated and non-automated traffic

In the previous chapter, we introduced the basic heterogeneous fleet sizing problem considering traffic congestion. In this chapter, building on the discussed problem, we examine interactions between centrally dispatched taxis and privately-owned human-driven vehicles. To model this, we propose a bi-level framework where the lower level captures mixed routing behaviour, and the upper level determines fleet sizes to maximise profit. A parallel genetic algorithm, embedded with a tailored algorithm for the lower level, is introduced. Numerical experiments on a small network and the Delft network in the Netherlands demonstrate the solution method's performance.

This chapter is structured as follows. Section 3.1 introduces the background information. Section 3.2 presents the literature review. Section 3.3 describes the mathematical model of the proposed bi-level framework. In Section 3.4, a detailed explanation of the solution methods is provided. Section 3.5 presents the case studies. Finally, conclusions and future outlook are given in Section 3.6.

## 3.1 Introduction

Uber's establishment in 2009 marked the beginning of the e-hailing industry. Since then, an increasing number of e-hailing services by the so-called Transportation Network Companies (TNCs), such as Uber, Lyft and Didi, have emerged globally, revolutionising urban mobility patterns and passenger travel behaviour (Liang et al., 2020). To maximise profit, a TNC must make a series of decisions, both at the planning level (fleet sizing, pricing strategy, service quality level) and the operational level (ride-matching and vehicle routing). Since transport demand and transportation infrastructure evolve through time, planning and operations must be adaptable to the existing situation at each point in time to obtain the highest performance.

Nowadays, e-hailing services are anticipating an upcoming revolution in urban mobility and road infrastructure that will result from the emergence of automated vehicles (AVs). AVs, which can be centrally controlled as "moving robots", are likely to be deployed by TNCs, promising to benefit service providers by eliminating both drivers' costs and their driving preferences (Ashkrof et al., 2022b) and offering continuous and high-quality door-to-door trip service Liang et al. (2020); Yang et al. (2020). Despite the great potential benefits, it is still impossible to convert all vehicles to AVs at once because of the high costs of fleet renewal and infrastructure adaptation. It is more realistic to expect in the near future that a small number of AVs are being used and that human-driven vehicles (HVs) gradually phase out. Throughout this transition period, AVs and conventional vehicles (CVs) will inevitably coexist in mixed traffic on the urban network (Chen et al., 2017). However, numerous studies have demonstrated that mixed traffic is less efficient than a fully automated traffic system (Yang et al., 2016; Olia et al., 2018). To improve traffic efficiency, many researchers envisioned that city planners and government agencies may have to dedicate specific traffic lanes (Chen et al., 2016; Liu & Song, 2019), or areas (Chen et al., 2017; Madadi et al., 2020; Conceição et al., 2021) to AVs. These areas, which we will designate in this chapter as AVs-only zones, will gradually expand until the entire road network is fully transformed into an automated and connected shared mobility system. For a TNC or a taxi company that wishes to modernise its services, decisions need to be taken adaptively and dynamically with the expansion of such areas.

Among all decisions, fleet sizing is one of the most critical determinants for a TNC as it determines the number of trips that can be satisfied and therefore the company's market share and associated profit. The literature on the fleet sizing problem is extensive. Recently, great interest has been rising in the heterogeneous fleet sizing problem under a mixed driving environment (Scherr et al., 2019; Yang et al., 2020; Mo et al.,

2022). Some consider this problem in a mixed driving environment with the emergence of AVs-only zones (Scherr et al., 2019) or mixed operation zones (Guo et al., 2021b). However, less attention has been devoted to dynamic interactions between road users and the infrastructure, resulting in endogenous traffic congestion. None of them considers the different routing behaviours among all road users.

The fleet sizing decision is dependent on the operational decisions of trip assignment and taxi routing. In a mixed driving environment, taxis' route choices are heavily influenced by privately owned human-driven vehicles (PVs). However, very few studies on fleet sizing problems have considered the impact of PVs' routing behaviour. Unlike taxis coordinated by a TNC to maximise system-wide profits, PVs behave selfishly, with drivers choosing routes that minimise their individual costs. These distinct routing behaviours align with the concepts of system optimum (SO) and user equilibrium (UE), respectively, in the traffic assignment theory (Sheffi, 1985). It is important to note that the "system" under examination in this chapter specifically pertains to the taxi system operated by the TNC, rather than the entire transportation system. To ensure realistic fleet sizing decisions, it is essential not to overlook the routing of PVs; this requires explicit modelling. The key challenge in this chapter is to integrate the different routing behaviours and the complex operational decisions of taxis in one model to determine a realistic optimal fleet size.

We propose a fleet sizing model for a TNC that deploys a heterogeneous fleet of both automated taxis (ATs) and conventional taxis (CTs) during a transition period while taking into account the dynamic interactions of this fleet with PVs and the road infrastructure. Along with the expansion of the AVs-only zone, the TNC needs to determine the optimal fleet size for ATs and adjust the current fleet size of CTs to better meet passengers' demand who can have a preference for using either ATs or CTs. Therefore, three types of traffic participants are considered in the model: ATs at level 5 automation (On-Road Automated Driving (ORAD) committee, 2021), CTs driven by taxi drivers and PVs driven by their owners. ATs at level 5 are capable of driving freely on the entire network, while HVs (CTs and PVs) are only allowed to drive outside the AVs-only zone. The exclusion of privately-owned AVs is motivated by two primary factors. Firstly, numerous researchers envision a future where AVs are mainly used through sharing and pooling options integrated into public transport, rather than being privately owned (Stoiber et al., 2019; Liang et al., 2020); secondly, we anticipate that the overall number of privately-owned AVs will likely be relatively small compared to the number of ATs. This projection is attributed to the expected high cost of AVs and the prevailing trend of favouring public transport and active modes of transport in cities, thereby limiting private vehicle ownership (Nieuwenhuijsen & Khreis, 2016; UITP, 2017).

To address the aforementioned problem and fill the gap in the current literature, we propose a bi-level framework to give managerial insights with regards to heterogeneous fleet sizing decisions (CTs and ATs) for a TNC along with the expansion of the AVs-only zone, also investigating the impacts of the AVs-only zone on traffic. At the upper level, the optimal fleet size of CTs and ATs is determined with the aim of maximising the profit of a TNC on the premise of fulfilling the travel demand. At the lower level, the dynamic routing interaction among travellers with UE (PVs) and SO (CTs, ATs) routing behaviours is captured. This behaviour will in turn have an impact on the decision-making process at the upper level. The traffic congestion effect is expressed through the dynamic travel times at the lower level.

The contributions of this chapter are summarised as follows:

- The studied problem enriches the well-investigated fleet sizing problem for on-demand mobility services by incorporating the following new elements: (1) infrastructure evolution: the emergence and expansion of AVs-only zones; (2) multiple players with different routing behaviour: PVs (following the UE) and centrally dispatched taxis (following the SO); (3) endogenous congestion caused by the routing of both the e-hailing taxis and PVs.

- We introduce a novel methodology that approximates the dynamic mixed equilibrium and integrates the comprehensive planning and operational decisions for taxis (fleet sizing, matching, routing, relocation, and parking) within a bi-level mixed-integer linear programming (MILP) model.

- We develop a tailored genetic algorithm framework to tackle the bi-level model. To solve the lower-level model, a two-stage solution framework is proposed. The first stage introduces a method for generating a path pool by determining the maximum allowable travel distances for all OD pairs, effectively constraining the path pool to a manageable size. In the second stage, using the path pool as input, we employ an iterative procedure embedded with a weight determination algorithm to compute the approximated mixed equilibrium model.

- This study provides TNCs as well as city planners and the government with managerial insights regarding the potential impact of AV-related infrastructure.

Given the nature of the proposed model as a MILP, a perfect mixed equilibrium cannot be guaranteed. We fully acknowledge that this is not a perfect model to capture the dynamic mix equilibrium, and we can only approximate the dynamic mixed equilibrium at a macroscopic level and ignore microscopic traffic dynamics. However, this research may provide insights into fleet management challenges, especially when considering the route choices made by PVs in a congested environment.

## 3.2 Literature review

### 3.2.1 Fleet sizing problem for e-hailing services

The problem we study is the extension of the well-known fleet sizing and mix vehicle routing problem (FSMVRP). Different from the typical fleet sizing problem, FSMVRP relaxes the assumption that all vehicles need to be homogeneous, which is more realistic in real-world applications. Heterogeneous fleet composition is considered but not limited to the following cases: vehicles with different capacities (Hiermann et al., 2016; Balac et al., 2020), vehicles with different cost structures (Hiermann et al., 2016), and vehicles with different functional types such as cars and buses (Santos & Correia, 2021). Including AVs in on-demand mobility brings non-negligible benefits which distinguish AV's cost structure from that of HVs, and may result in potential cost savings. This boosts the need to investigate the fleet sizing problem once AVs enter the market.

Research has demonstrated the need to investigate the heterogeneous fleet sizing problem on shared mobility deploying both AVs and HVs in a mixed driving environment. Mo et al. (2022) stated that managerial decisions such as fleet size and pricing for AVs and HVs need to be determined properly and attention needs to be paid to the trade-off between these two types of services. To this end, they proposed an aggregated market model to examine how fleet sizing and pricing decisions for both types of services affect the demand rates, riders' utility, and riders' waiting time with congestion effects. Based on the numerical analysis, they suggested that more AVs should be arranged than HVs even under the scenario where AVs had a higher depreciation cost.

However, few studies consider this problem together with the emergence of specific intelligent infrastructure. More recently, Guo et al. (2021b) foresaw the emergence of the mixed operation zone (MOZ), an urban zone in which AVs and HVs can operate together. Based on the emergence of MOZ, they conducted research to determine the robust minimum fleet size of AVs and HVs deployed by on-demand rides services, taking demand uncertainty into account, and investigating the impacts of this zone on the performance of the service. A two-stage robust optimisation model is proposed and solved optimally. The objective function of this model is to minimise the total number of vehicles required to fulfil the travel demand. However, the minimum fleet size to serve all the demand is not necessarily the optimal fleet size for the on-demand mobility system as the minimum fleet may not lead to the greatest profit. For instance, a small fleet is likely to result in a longer detour distance (Militão & Tirachini, 2021), which might cause high operational costs. As a profit-oriented company, a TNC would rather systematically make the fleet sizing decision by analysing various factors, such as the total operational cost, the depreciation cost, the salaries paid to drivers, and the

congestion effect caused by the fleets, etc. Nevertheless, it is worthwhile to investigate the relationship between the minimum and optimal fleet size, as well as the trade-off between fleet sizes of different vehicle types. Fan et al. (2022) examined how the gradual expansion of the AVs-only zone affects fleet size decisions during the transition period from a conventional to a fully intelligent road network. They envisioned two business models for on-demand mobility services and included endogenous traffic congestion in the model. However, they did not take into account the distinct routing behaviours of AVs and HVs, which will be the focus of this chapter.

Mainly three types of modelling techniques have been used to tackle fleet sizing problems: simulation-based techniques (Fagnant & Kockelman, 2018; Yi & Smart, 2021; Wang et al., 2022a), optimisation-based techniques (Allahviranloo & Chow, 2019; Balac et al., 2020; Guo et al., 2021b), and hybrid methods combining the two (Militão & Tirachini, 2021). Simulation-based techniques can reproduce complex scenarios by considering the diverse behaviours of road users and monitoring their dynamic interactions. However, they are usually time-consuming because a large number of simulations with varying fleet sizes are required to evaluate the system's performance. When various fleet types are considered, the number of possible combinations could be very high. Moreover reproducing realistic route choices of a mixed fleet of vehicles also takes time in a simulation-based methodology.

Among the optimisation-based techniques, fleet sizing problems are typically modelled as a single-level mixed integer linear programming model (Koç et al., 2016; Balac et al., 2020; Santos & Correia, 2021), or a bi-level model (Allahviranloo & Chow, 2019), solved by exact methods (Balac et al., 2020; Santos & Correia, 2021; Fan et al., 2022), or heuristic methods (Renaud & Boctor, 2002; Brandão, 2009; Koç et al., 2016), or hybrid methods (Wang et al., 2019). For some simple scenarios, a single-level model is sufficient when minimising the fleet size is the only goal. Another typical scenario is when all vehicles are under the control of a central agent (eg. TNC, or government). In this case, the fleet size decisions together with the route choice of vehicles are taken over by the operator.

For a more complex problem involving interactions between the supply strategies of the fleet operators and the route choices or activity schedule of all travellers (not just the deployed fleets) in the road network, a bi-level model is required. This type of problem is known as the network design problem. At the upper level, operators make profit-maximising decisions. Travellers respond to those decisions at the lower level. Allahviranloo & Chow (2019) studied the fleet sizing problem in a future scenario in which users of autonomous transport services may share ownership of AVs and pay for the time slots for daily activities. A bi-level model was formulated. At the lower level, demand was determined by the activity scheduling decisions. This decision was

in turn influenced by the fleet capacity and the time slot prices determined at the upper level. Li & Liao (2020) proposed a bi-level framework for the network design problem to investigate the optimal deployment of shared AVs (SAVs). The optimal SAV hub locations, fleet size and the initial distribution of SAVs were determined at the upper level. Based on these decisions, the activity-travel scheduling was modelled at the lower level. When modelling the interactions between AVs and CVs, some researchers use a leader-follower game structure, in which AVs are the leaders and HVs are the followers. In this system, AVs are centrally controlled by the operators and CVs respond to the coordination of AVs (Yang et al., 2020).

As a complement to the existing literature, this chapter aims to investigate the interactions between the operator's strategy and travellers' behaviour in the context of the emergence of AVs-only zones. This type of problem is best characterised by a bi-level framework. At the lower level, the route choices of taxis and PVs are modelled, which follow the SO and UE principles, respectively. At the upper level, fleet sizing decisions are made to maximise profit. If we disregard the flow of PVs, all decisions (fleet size, number of served trips, route choices of taxis) can be made at the same level, according to the SO principle.

### 3.2.2 Vehicle routing problem (VRP) and Traffic assignment (TA)

As stated previously, the problem we study is an extension of the FSMVRP, which is further integrated with important TA concepts. These two fields share non-negligible similarities but also have distinct features. In a traditional VRP, the optimal routes of a fleet of vehicles are determined to traverse the road network from one depot to another to deliver and/or pick up a set of goods/customers (Laporte, 2009). In the context of on-demand mobility transport, a few decisions must be made, including trip assignment, passenger pick-up and delivery process, vacant vehicles' relocation and parking decisions, under the restrictions of time windows and vehicle capacity. Based on these decisions, more managerial strategies/decisions of the fleet operator could be included in the model, such as fleet size, pricing, service quality, etc. The dynamic traffic assignment (DTA) models traffic flow between a specific origin and destination pair without considering the planning and operational decision-making process (order dispatching, vehicle parking, vehicle relocation, etc) in the context of on-demand mobility services. Nevertheless, TA can capture the congestion effect incurred by the interactions between vehicles and infrastructures, as well as modelling the different routing behaviours of travellers. The methodology proposed in this chapter will bridge these two research fields by modelling the congestion effects and different routing behaviours of travellers within an FSMVRP.

A few researchers have attempted to bridge the VRP with the TA. Correia & Van Arem (2016) proposed a successive average framework to solve the dynamic user optimum privately-owned AV assignment. However, rather than directly assigning the flow to the minimum cost path on the network, the routing and parking decisions of a household's AV are determined by solving a proposed MILP model to minimise the total generalised cost of transporting a single household. The congestion effect is captured by the flow-dependent link travel time, which will be updated outside the MILP model using a non-linear Bureau of Public Roads (BPR) function. Van Essen & Correia (2019) proposed a novel exact formulation to approximate the dynamic user optimum by incorporating it into a MILP model. The objective of the model is to minimise the maximum relative deviation from the minimum cost for each household. By doing so, households will have similar relative deviations. The traffic congestion effect described by the non-linear BPR function is involved in the model in a linear form. Liang et al. (2018) introduced an optimisation model for trip assignment and dynamic routing of ATs to maximise the total profit of the operator. To describe the congestion level of each link, they used breakpoints on a BPR function while embedding it in the proposed MILP model. Chen & Levin (2019) claimed that dynamic UE assignment is more promising for on-demand mobility services, because of the competition among mobility service providers. They firstly developed a static UE TA model for the route choice of AVs between urban origins and destinations. Based on the solution, a linear programming model is solved to specify the optimal rebalancing flow. This static model is converted into a dynamic one by adding the time dimension. Liu et al. (2020) considered an ideal scenario where all the vehicles operate with the SO principle. They firstly proposed a vehicle-based arc-based integer programming model in the space-time-state network which is similar to the VRP problem. Then, based on the generated mapping information of vehicle-passenger and vehicle-arc, they further developed a flow-based path-based linear programming model from the perspective of DTA and solved it by a column-pool-based approximation method.

A challenge for our problem is to model the dynamic mixed equilibrium considering both SO and UE principles in an FSMVRP which is usually a MILP model. Related works on modelling the mixed equilibrium in TA are mostly focused on static scenarios (Bagloee et al., 2017; Chen et al., 2017; Zhang & Nie, 2018; Kashmiri & Lo, 2022; Zhang et al., 2022; Ke & Qian, 2023), day-to-day dynamic systems (Li et al., 2018; Liang et al., 2023), and dynamic scenarios (Guo et al., 2021a; Mansourianfar et al., 2021, 2022; Hoang et al., 2023), but ignore the detailed vehicle operations (relocation and parking), trip assignment and vehicle dispatching, and the managerial decisions from the perspective of a TNC. To overcome these shortcomings, in this chapter, we consider the feedback of operational strategies of taxis on the network

traffic conditions and propose a bi-level framework to determine the planning and operational decisions while approximating the dynamic mixed equilibrium in a typical working day. Our work shares a few similarities with the study by Ge et al. (2021), which proposes an SAV matching and routing problem in a traffic assignment context, considering the endogenous traffic congestion from both CVs and SAVs. In their approach, a bi-level programming model is developed with SAVs as leaders and CVs as followers. Although this problem is investigated under a static setting, they suggest the possibility of extending the model to dynamic traffic conditions. Compared with the referred work, our study aims to determine the optimal planning decisions while also providing more detailed operational decision chains, including detailed parking choices, relocation decisions from trip to trip, and endogenous congestion caused by all the road users under dynamic traffic settings. To the best of our knowledge, the FSMVRP considering traffic congestion and the approximated mixed equilibrium has rarely been studied in the context of on-demand mobility services.

## 3.3 Problem formulation

The proposed bi-level framework is presented in this section as a bi-level MILP model. In Section 3.3.1, we first introduce the problem. Then, we propose the mathematical formulation of the upper level and the lower level in Sections 3.3.2 and 3.3.3, respectively.

### 3.3.1 Problem description and modelling framework

The demand of travellers heading from origins to destinations triggers the need to plan the operation of e-hailing services and vehicle movements on the road network. The model structure that is supposed to solve the problem is presented in Figure 3.1 depicting the decisions, elements (e.g. demand, game players, and infrastructure) and their relations.

In terms of planning, we assume that the demand for the optimisation period is known in advance and the overall travel demand in an urban area is fixed for a given optimisation period. This assumption makes sense for a planning problem that this study addresses. The overall travel demand is divided into two groups: those who drive their own vehicles, and those who choose to ride in taxis. For the first group of travellers, driving their PVs will always be the preferred mode of transportation, unless the destination is inaccessible to HVs due to the restrictions imposed by the AVs-only zone. These travellers will then have to switch to ATs. No choice modelling is involved

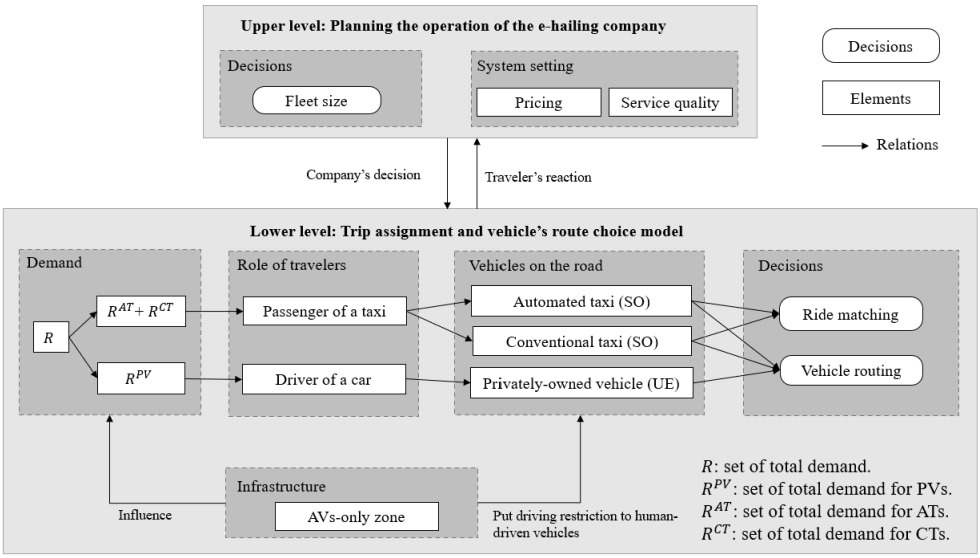



*Figure 3.1: Decisions, elements and their relations in the bi-level optimisation problem.*

because it is not the focus of our problem. In a future study, when analysing the effect of AVs-only zones on travellers' behaviour, choice modelling can be incorporated.

The demand for different types of taxis is determined by customers' preferences, which are known in advance. This means that travellers can choose the vehicle type by themselves in case the trip can be served by either type of taxi. Considering travellers' preferences will significantly increase users' satisfaction with the e-hailing service. Assuming that travellers who use e-hailing taxi services are fully aware of the services provided by the TNC and the available options of the vehicle types, they will adapt their behaviour to the on-demand mobility system and make feasible trips through the app-based service provider platform. Above a minimum service rate to guarantee service quality, the company will serve those trips that generate the most profit. Once the trip is rejected by the system, the traveller will opt for public transit, such as bus, subway, or train, which are not included in our model as they barely contribute to the congestion on the road network.

The movement of passengers and vehicles is aggregated into flows in the model if their trips have the same origin, destination and departure time. This avoids tracking each vehicle independently, thereby reducing the number of decision variables. On the roads, vehicle flows composed of PVs, CTs and ATs make route choices and then contribute to congestion. Congestion is quantified by the dynamic link travel time as

a function of traffic flow. The varying link travel time will, in turn, affect the route choices of the vehicles. The interplay between the route choice of the vehicles and dynamic travel time considering traffic congestion is also considered in this model. Despite treating the vehicle movements as flows, vehicles in the same group are allowed to take different routes and have different arrival times at the destination to balance the network usage.

A time-space network is used to capture the dynamic interactions among road users. This network is defined by duplicating the directed physical network $(N, L)$ at each time instant $t \in T$, where $N$ and $L$ denote the set of nodes and road links. On the time-space network, vehicles move on links $(i_{t_1}, j_{t_2}) \in G$, indicating the flow movement from node $i \in N$ to node $j \in N$ from time instant $t_1 \in T$ to time instant $t_2 \in T$. To specify the driving area of different types of vehicles $m \in M$, extra sets are introduced as $N^m$ and $G^m$ to denote the nodes and links in the time-space network that can be used by the vehicles of type $m \in M$. By doing so, the driving restrictions for different types of vehicles are easily included. In our problem, each type of vehicle has a corresponding driving area: CTs and PVs are not permitted to use the links inside the AVs-only zone; ATs of level 5 automation, on the other hand, can drive everywhere on the urban network. The proposed model can easily be extended to a more general situation involving additional vehicle types such as level 4 AVs that can only circulate in certain areas. We assume that taxis are only permitted to park at designated nodes that are identified as TNC's parking depots. The parking depots that are accessible to taxis of type $m \in \{CT, AT\}$ are designated as $N_p^m$.

Given the driving restrictions imposed on HVs, the unique TNC operator in the city will assign the appropriate type of vehicle to fulfil the incoming trip requests. There are three types of trips regarding the location of the origin and destination (shown in Table 3.1).

*Table 3.1: Type of trips and serving vehicles*

| Demand | Origin | Destination | CT | AT |
|---|---|---|---|---|
| Type 1 | AVs-only zone | AVs-only zone | | ✓ |
| Type 2 | Outside the AVs-only zone | Outside the AVs-only zone | ✓ | ✓ |
| Type 3 | AVs-only zone | Outside the AVs-only zone | | ✓ |
| | Outside the AVs-only zone | AVs-only zone | | ✓ |

Moreover, several assumptions are made underlying the proposed modelling framework: (1) No vehicles are allowed to go back to a previously visited arc in the road network when heading from the origin to the destination of a trip; (2) The origin and destination node of a group of trips will be visited only once while delivering the

clients; (3) No ride-pooling is considered in this study. Each vehicle is limited to carrying a single passenger at a time. (4) The capacity of links within the AVs-only zone is larger than the capacity of the links outside the AVs-only zone which is to represent the added traffic efficiency of these vehicles (Chen et al., 2017; Madadi et al., 2020).

The following sections introduce the mathematical formulation of the bi-level MILP model. The notation used in this model is presented in Table 2.

*Table 3.2: Notation*

| Notation | Description |
|---|---|
| **Sets** | |
| $M$ | $= \{CT, AT, PV\}$, set of vehicle types. |
| $T$ | $= \{0, \ldots, t, \ldots, s\}$, set of time instants in the operation period. |
| $N$ | $= \{1, \ldots, i, \ldots\}$, set of nodes. |
| $L$ | $= \{\ldots, (i, j), \ldots\}$, set of road links between nodes in set $N$. |
| $G$ | $= \{\ldots, (i_{t_1}, j_{t_2}), \ldots\}$, set of links in the time-space network. |
| $R^m$ | $= \{1, \ldots, r, \ldots\}$, set of groups of trips served by vehicles of type $m \in M$, where each group of requests $r \in R^m$ has the same origin, destination, desired departure time, and latest arrival time at the destination. |
| $N^m$ | $\subseteq N$, set of nodes that can be used by vehicles of type $m \in M$. CTs and PVs can use the nodes outside the AVs-only zone and the nodes located at the border of the AVs-only zone; ATs can use all the nodes. |
| $N_P^m$ | $\subseteq N^m$, set of nodes allowing parking for taxis of type $m \in \{CT, AT\}$. |
| $G^m$ | $\subseteq G$, set of links that can be used by vehicles of type $m \in M$ in the time-space network. |
| $\Pi^r$ | $= \{1, \ldots, \pi, \ldots\}$, set of paths of group of trips $r \in R^{PV}$. |
| **Parameters** | |
| $p^0$ | Initial base fare in euros for using the taxis. |
| $p^m$ | Price per kilometre in euros/km for using a taxi of type $m \in \{CT, AT\}$. |
| $co^m$ | Unit driving operational cost in euros/km for vehicle type $m \in M$. |
| $cp$ | Salary of a driver in euros/time step. |
| $cd$ | Penalty for drop-off delay of passengers in euros/time step. |
| $cf^m$ | Depreciation cost in euros/vehicle in one hour for using vehicle type $m \in \{CT, AT\}$. |
| $ct$ | Perceived value of time cost for passengers driving PVs in euros/time step. |
| $s$ | Total number of time instants in the operation period. |
| $\alpha$ | Minimum service rate for orders. |
| $lb^m$, $ub^m$ | Lower bound and upper bound of taxi's fleet size of type $m \in \{CT, AT\}$. |
| $\omega$ | Calibrated weighting coefficient to combine two objective functions into one. |
| $\lambda$ | Predefined weighting coefficient to give priority to a certain term in the objective function. |
| $l_{ij}$ | Length of road link $(i, j) \in L$. |
| $Q_{ij}$ | Capacity of road link $(i, j) \in L$ in vehicles per time step. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of road link $(i, j) \in L$ in vehicles that fit on the road link from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |
| $t_{ij}^{\max}$ | Maximum travel time on road link $(i, j) \in L$. |

| | |
|---|---|
| $t_{ij}^{\min}$ | Minimum travel time on road link $(i,j) \in L$. |
| $o^r$ | Origin node for group of trips $r \in R^m, m \in M$. |
| $d^r$ | Destination node for group of trips $r \in R^m, m \in M$. |
| $a^r$ | Desired departure time for group of trips $r \in R^m, m \in M$. |
| $b^r$ | Latest arrival time for group of trips $r \in R^m, m \in M$. |
| $sd^r$ | Shortest travel distance for group of trips $r \in R^m, m \in M$. |
| $st^r$ | Shortest travel time assuming free-flow speed for group of trips $r \in R^m, m \in M$. |
| $n^r$ | Total number of trips for group $r \in R^m, m \in M$. |
| $D^{r\pi}$ | The length of the path $\pi \in \Pi^r$ used by trips in group $r \in R^{PV}$ |
| $M^r$ | Minimum travel cost for trips in group $r \in R^{PV}$. |
| $\delta_{ij}^{r\pi}$ | Incidence between road link $(i,j) \in L^{PV}$ and path $\pi \in \Pi^r$ in group of trips $r \in R^{PV}$, 1 if the link is part of the path; 0 otherwise. |

**Decision variables**

| | |
|---|---|
| $P^r$ | Integer variable representing the total number of served trips from group $r$, where $r \in R^m, m \in \{CT, AT\}$. |
| $PF^r_{i_{t_1} j_{t_2}}$ | Integer variable representing the passenger flow in the group of trips $r \in R^m$ served by vehicle type $m \in M$ in road link $(i,j)$, from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G^m, a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| $PF^{r\pi}_{i_{t_1} j_{t_2}}$ | Continuous variable representing the passenger flow of the group of trips $r \in R^{PV}$ using path $\pi \in \Pi^r$ that travels in road link $(i,j)$ from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2})$ where $\delta_{ij}^{r\pi} = 1$, $a^r \leq t_1 < t_2 \leq b^r$. |
| $V^m$ | Integer variable representing the taxi fleet size of type $m \in \{CT, AT\}$. |
| $E^{rt}$ | Integer variable representing the total number of passengers in group of trips $r \in R^m$ for vehicle type $m \in \{CT, AT\}$ arriving at time $t \in T$. |
| $F^m_{i_{t_1} j_{t_2}}$ | Continuous variable representing the vehicle flow of type $m \in M$ in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G^m$. |
| $W^m_{i_t}$ | Continuous variable representing the total number of taxis of type $m \in \{CT, AT\}$ parking at node $i \in N_P^m$ from time instant $t$ to $t+1$, with $t \in T$. |
| $K^{r\pi}$ | Continuous variable representing the generalised cost of trips in group $r \in R^{PV}$ using path $\pi \in \Pi^r$. |
| $K^r$ | Continuous variable representing the maximum general cost of trips in group $r \in R^{PV}$. |
| $F^{r\pi}$ | Integer variable representing the vehicle flow using path $\pi \in \Pi^r$ of group of trips $r \in R^{PV}$. |
| $A_t^{r\pi}$ | Binary variable which is 1 when at least one trip in group $r \in R^{PV}$ using path $\pi \in \Pi^r$ arrives at time $t \in T$, and 0 otherwise. |
| $X_{i_{t_1} j_{t_2}}$ | Binary variable which is 1 when any vehicle travels in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. |

### 3.3.2 Upper-level model (ULM): Planning for the TNC

The upper-level optimisation model denoted as [ULM] has the following mathematical formulation. The objective function is:

$$\begin{aligned}\textbf{[ULM]}\max Z = & \sum_{m\in\{CT,AT\}}\sum_{r\in R^m}\left(p^0P^r + p^mP^r sd^r\right) - s\cdot cp\cdot V^{CT} - \sum_{m\in\{CT,AT\}} cf^m V^m \\ & - \sum_{m\in\{CT,AT\}} co^m\left(\sum_{\left(i_{t_1},j_{t_2}\right)\in G^m} l_{ij}F^m_{i_{t_1}j_{t_2}}\right) \\ & - cd\sum_{m\in\{CT,AT\}}\sum_{r\in R^m}\left(\sum_{t\in T} tE^{rt} - a^r n^r - st^r n^r\right)\end{aligned} \tag{3.1}$$

Subject to:

$$E^{rt}, F^m_{i_{t_1}j_{t_2}} \in \arg\min\{\text{Objective function (3.5)-(3.7)} : \text{Constraints (3.8)-(3.28)}\} \tag{3.2}$$

$$lb^m \le V^m \le ub^m, \qquad \forall m \in \{CT,AT\} \tag{3.3}$$

$$\alpha n^r \le P^r \le n^r, \qquad \forall r \in R^m, m \in \{CT,AT\} \tag{3.4}$$

The upper-level objective function denoted as $Z$ is to maximise the total profit of the TNC. The first term represents the taxi fares paid by the passengers. Two types of fares are included: an initial fixed base fare $p^0$ once the order is accepted, and an additional price $p^m$ based on the shortest travel distance $sd^r$ of the trip $r \in R^m$ where $m \in \{CT,AT\}$. Here, the shortest travel distance is used rather than the taxis' actual travel distance in order to avoid taxis detouring and charging passengers more money. The second term represents the salaries paid to human drivers of the CT fleet. The third term defines the depreciation cost of the different types of taxis in the system. The depreciation cost of a vehicle of type $m$ represented by $cf^m$, is calculated as the vehicle's purchase price divided by its service life span. Both the second and the third terms describe the cost associated with the fleet size. The fourth term is the operation cost of vehicles on the entire network including fuels, maintenance and assurance costs. This is calculated by the total travel distance for all the taxis multiplied by the operational cost per unit denoted by $co^m$. The final term is the penalty for the drop-off delay of the client which is calculated by multiplying the delay cost $cd$ by the delay time. The delay time is calculated as the time difference between the passengers' actual riding time and the shortest travel time in free-flow speed.

In this upper-level model, the values of variables $F^m_{i_{t_1}j_{t_2}}$ and $E^{rt}$ are determined in the lower-level problem, as indicated in Equation (3.2). Constraints (3.3) impose an

upper bound and lower bound on the total fleet size of CTs and ATs which is explained in Section 3.4.2. Constraints (3.4) guarantee that the number of trips served in the group of trips $r \in R^m$ should be less than the group's demand, but greater than the minimum number required to ensure service quality.

### 3.3.3 Lower-level model (LLM): Mixed routing model for taxis and PVs

For the lower-level problem, we describe the routing behaviour of heterogeneous traffic participants within a MILP model. Unlike the traditional TA problem, our methodology tackles a discrete optimisation problem within a time-space network framework rather than a continuous optimisation problem. This allows us to model both planning and operational decisions, whilst still capturing the impact of varying congestion resulting from the routing of the vehicles. In our problem formulation, integer variables are used to represent link travel times and passenger flows. However, due to the inherent nature of the integrality of time and flow, it becomes infeasible to achieve the traditional UE where travellers in all paths for a given O-D pair experience equal travel costs. This integrality aspect poses a challenge when trying to directly impose UE constraints in the MILP framework. Alternatively, brought from Van Essen & Correia (2019) the concept of approximated dynamic UE in mathematical programming, we propose a new method to approximate the mixed equilibrium (both UE and SO) in a MILP model.

The approximated mixed equilibrium used in this chapter is realised by the following steps. Firstly, in a dynamic setting we approximate the UE by minimising the difference between the cost of all routes for the same O-D pairs. This is accomplished by initially minimising the maximum relative deviation from the minimum cost and then minimising the total costs of PVs so that the costs of all the used paths have similar relative deviations. Secondly, when modelling the SO, the "system" we target is the TNC rather than the entire transportation system. The objective is to minimise the overall cost of taxi routing by optimally assigning clients to taxis and determining taxis' route choices. Subsequently, we approximate the mixed equilibrium by formulating a bi-objective optimisation model that considers the two independent objectives of taxis and PVs. We further propose an approach to balance the contribution of these two objectives.

In terms of modelling bi-objective optimisation problems, one of the most extensively used classic techniques is the weighted-sum method, which can convert the two objective functions into one by using a weighting coefficient. The weighting coefficient indicates the decision maker's preference or the relative importance of the two objec-

tives. Thus, it is critical to properly assign it a value. In the mixed routing problem, when the network is congested and the objectives of all road users cannot be satisfied simultaneously, vehicles with different routing objectives are usually competing for the best routes. Nonetheless, the objective functions of taxis and PVs should be given the same priority. Thus, the weighting coefficient should balance the contribution of the two objective function values. An iterative weight determination method is proposed to produce the desired traffic patterns on the network. A detailed description of this method can be found in Appendix 3.A.1.

The route choices of the taxis and PVs are modelled differently. Assume that the PVs consider generalised costs as the routing criteria, which contain a travel time-related cost and a distance-related cost. When modelling the routing behaviour of PVs, it is necessary to compare the generalised travel costs of different paths for the same O-D pair. To specify the travel time and distance associated with a particular path, path-based variables will be required to describe the movement of the passengers. For taxis, no paths will be compared when modelling their route choices because one is aiming for the system optimal flow distribution. As a result, arc-based variables are enough to describe the taxi flow.

Path sets containing alternatives for a given O-D pair will be generated before the optimisation. Some restrictions are taken into account when generating paths: first, the shortest travel time of using a path should be within the time window indicated by passengers which is the latest arrival time minus the departure time; paths with repeated arcs are not included as we assume that vehicles will not detour back to a previously visited arc in a directed network when heading from the origin to the destination due to the significantly increased travel distance cost. Even so, enumerating all the paths with the proposed restrictions in an urban scale network is still unrealistic as the huge number of paths could significantly increase the scale of decision variables, leading to a computational burden. Section 3.4.1 describes how to find small-scale path sets that include paths that PVs will take.

We formulate the described LLM as follows:

**Objective function**

$$\textbf{[LLM]} \qquad \min J = \omega \cdot J^T + (1-\omega) \cdot J^P \tag{3.5}$$

where

$$J^T = \sum_{m \in \{CT,AT\}} co^m \cdot \left( \sum_{(i_{t_1}, j_{t_2}) \in G^m} F^m_{i_{t_1} j_{t_2}} \cdot l_{ij} \right) + cd \cdot \sum_{r \in R^m} \sum_{m \in \{CT,AT\}} \left( \sum_{t \in T} E^{rt} \cdot t \right) - a^r \cdot P^r - st^r \cdot P^r \tag{3.6}$$

$$J^P = \lambda \cdot \sum_{r \in R^{PV}} \frac{K^r}{M^r} + \sum_{\pi \in \Pi^r, r \in R^{PV}} \sum_{\left(i_{t_1}, d_t^r\right) \in G^{PV}} PF^{r\pi}_{i_{t_1} d_t^r} \left(D^{r\pi} \cdot co^{PV} + (t - a^r) \cdot ct\right) \quad (3.7)$$

Taxis have an objective function denoted by $J^T$ that seeks to minimise the total operational costs and the drop-off delay penalty of the clients. PVs have an objective function denoted as $J^P$ that minimises first the maximum generalised travel cost $K^r$ relative to the lowest possible generalised travel cost $M^r$ for all groups of trips $r \in R^{PV}$. Additionally, it seeks to minimise the total generalised cost across all trips, taking into account that costs with a lower relative deviation than the maximum relative deviation can also be minimised. To prioritise the first term of the objective function, which aims to minimise the cost difference between routes for the same OD pair, we introduce a weighting coefficient $\lambda$ that gives absolute priority to this term. A detailed description of how to determine the value of $\lambda$ can be found in Appendix 3.A.2. As previously stated, we use the weighted-sum method to combine $J^T$ and $J^P$ into one single objective function (weight $\omega$). The objective function is constrained by the following:

**Constraints for taxis:**

$$P^r = \sum_{j_t \mid \left(o^r_{a^r}, j_t\right) \in G^m} PF^r_{o^r_{a^r} j_t}, \qquad \forall r \in R^m, m \in \{CT, AT\} \quad (3.8)$$

$$P^r = \sum_{t \in T \mid a^r + st^r \leq t \leq b^r} E^{rt}, \qquad \forall r \in R^m, m \in \{CT, AT\} \quad (3.9)$$

$$E^{rt} = \sum_{i_{t_1} \mid \left(i_{t_1}, d_t^r\right) \in G^m} PF^r_{i_{t_1} d_t^r}, \qquad \forall r \in R^m, m \in \{CT, AT\}, t \in T \quad (3.10)$$

$$\sum_{j_{t_2} \mid \left(d^r_{t_1}, j_{t_2}\right) \in G^m} PF^r_{d^r_{t_1} j_{t_2}} = 0, \qquad \forall r \in R^m, m \in \{CT, AT\}, t_1 \in T, a^r + stt^r \leq t_1 \leq b^r \quad (3.11)$$

$$\sum_{i_{t_1} \mid \left(i_{t_1}, o^r_{t_2}\right) \in G^m} PF^r_{i_{t_1} o^r_{t_2}} = 0, \qquad \forall r \in R^m, m \in \{CT, AT\}, t_2 \in T, a^r \leq t_2 \leq b^r \quad (3.12)$$

$$\sum_{j_{t_0} \mid \left(j_{t_0}, i_{t_1}\right) \in G^m} PF^r_{j_{t_0} i_{t_1}} = \sum_{j_{t_2} \mid \left(i_{t_1}, j_{t_2}\right) \in G^m} PF^r_{i_{t_1} j_{t_2}}, \qquad \forall r \in R^m, m \in \{CT, AT\}, t_1 \in T, a^r < t_1 < b^r, i \in N^m, i \neq o^r, i \neq d^r \quad (3.13)$$

$$\sum_{r \in R^m} PF^r_{i_{t_1} j_{t_2}} \leq F^m_{i_{t_1} j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G^m, m \in \{CT, AT\} \quad (3.14)$$

$$\sum_{(i_0, j_t) \in G^m} F^m_{i_0 j_t} + \sum_{i \in N^m_P} W^m_{i_0} = V^m, \qquad \forall m \in \{CT, AT\} \quad (3.15)$$

$$\sum_{j_{t_1}|(j_{t_1},i_t)\in G^m,t_1<t} F^m_{j_{t_1}i_t} + W^m_{i_{t-1}} = \sum_{j_{t_2}|(i_t,j_{t_2})\in G^m,t<t_2} F^m_{i_t j_{t_2}} + W^m_{i_t},$$
$$\forall t \in T, 0 < t < s, i \in N^m_P, m \in \{CT, AT\} \quad (3.16)$$

$$\sum_{j_{t_1}|(j_{t_1},i_t)\in G^m,t_1<t} F^m_{j_{t_1}i_t} = \sum_{j_{t_2}|(i_t,j_{t_2})\in G^m,t<t_2} F^m_{i_t j_{t_2}},$$
$$\forall t \in T, 0 < t < s, i \in N^m \setminus N^m_P, m \in \{CT, AT\} \quad (3.17)$$

Taxis serving the trips in the same group $r \in R^m$ depart from the origin $o^r$ at the same time, but are permitted to take different routes and arrive at the destination at different times. Constraints (3.8)-(3.10) ensure that passenger flows departing from node $o^r$ at time $a^r$ and arriving at the destination node $d^r$ are equal to the total number of trips served in group $r \in R^m$. Constraints (3.11) and (3.12) guarantee that the passenger flows start at the origin node and end at the destination node. Constraints (3.13) define the conservation of passenger flow through intermediate nodes of the network. Then, the passenger flows and the vehicle flows are linked via constraints (3.14), which make sure that the total number of passengers travelling on road link $(i, j)$ from time instant $t_1$ to time instant $t_2$ will never exceed the total number of taxis on the same link. Given the fleet size of CTs and ATs, constraints (3.15) guarantee that the total number of taxis circulating on road link $(i, j)$ or parking at depot $i \in N^m_p$ at the start of the service period is consistent with the fleet size specified. In this case, the fleet sizes $V^m$ of taxis of type $m$ are exogenous variables, whose values are determined at the upper level. The vehicle flow equilibrium for nodes that allow or not allow vehicle parking is defined by constraints (3.16) and (3.17) respectively.

**Constraints for PVs:**

$$\sum_{\pi\in\Pi^r} F^{r\pi} = n^r, \qquad \forall r \in R^{PV} \quad (3.18)$$

$$F^{r\pi} = \sum_{j_{t_2}|(o^r_{a^r},j_{t_2})\in G^{PV},\delta^{r\pi}_{o^r j}=1} PF^{r\pi}_{o^r_{a^r},j_{t_2}}, \qquad \forall \pi \in \Pi^r, r \in R^{PV} \quad (3.19)$$

$$F^{r\pi} = \sum_{(i_{t_1},d^r_{t_2})\in G^{PV},\delta^{r\pi}_{id^r}=1} PF^{r\pi}_{i_{t_1},d^r_{t_2}}, \qquad \forall \pi \in \Pi^r, r \in R^{PV} \quad (3.20)$$

$$\sum_{j_{t_0}|(j_{t_0},i_{t_1})\in G^{PV},\delta^{r\pi}_{ji}=1} PF^{r\pi}_{j_{t_0}i_{t_1}} = \sum_{j_{t_2}|(i_{t_1},j_{t_2})\in G^{PV},\delta^{r\pi}_{ij}=1} PF^{r\pi}_{i_{t_1},j_{t_2}}, \qquad \forall \pi \in \Pi^r, r \in R^{PV}, t_1 \in T,$$
$$a^r < t_1 < b^r, i \in N^{PV}, i \neq o^r, i \neq d^r \quad (3.21)$$

$$F^{PV}_{i_{t_1} j_{t_2}} = \sum_{\pi \in \Pi^r, r \in R^{PV}} PF^{r\pi}_{i_{t_1} j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G^{PV} \tag{3.22}$$

$$A^{r\pi}_t \geq \frac{\sum_{i_{t_1} | \left(i_{t_1}, d^r_t\right) \in G^{PV}} PF^{r\pi}_{i_{t_1}, d^r_t}}{n^r}, \qquad \forall \pi \in \Pi^r, r \in R^{PV}, t \in T \tag{3.23}$$

$$K^{r\pi} = \sum_{t \in T} A^{r\pi}_t \left( D^{r\pi} \cdot co^{PV} + (t - a^r) \cdot ct \right), \qquad \forall \pi \in \Pi^r, r \in R^{PV} \tag{3.24}$$

$$K^r \geq K^{r\pi}, \qquad \forall \pi \in \Pi^r, r \in R^{PV} \tag{3.25}$$

Constraints (3.18) ensure that the total number of trips using different paths $\pi \in \Pi^r$ in group $r \in R^{PV}$ equals the total number of trips in group $r \in R^{PV}$. If link $(i, j) \in L^{PV}$ belongs to path $\pi \in \Pi^r$ of group of trips $r \in R^{PV}$, the link flow for this path should equal the path flow, as indicated in constraints (3.19) and (3.20). Constraints (3.21) describe the passenger flow conservation for trips in group $r \in R^{PV}$ using different paths $\pi \in \Pi^r$ at all nodes excluding their origin and destination node. Constraints (3.22) link the passenger flow to the vehicle flow. To compare the generalised cost of all the used paths, we have to calculate the path lengths and their corresponding travel times. The length of the path $\pi \in \Pi^r$ in group of trips $r \in R^{PV}$ is calculated as the sum of length of link $(i, j) \in L^{PV}$ if link $(i, j)$ is part of the path, which is $D^{r\pi} = \sum_{(i,j) \in L^{PV}} l_{ij} \cdot \delta^{r\pi}_{ij}$. Constraints (3.23) determine whether PVs in the group of trips $r$ using the path $\pi \in \Pi^r$ arrive at the destination at time instant $t \in T$. Then, the generalised cost of using path $\pi \in \Pi^r$ for group of trips $r \in R^{PV}$ is calculated as expressed by constraints (3.24). Knowing the costs of all the used paths from group of trips $r \in R^{PV}$, the maximum cost over all the trips is determined by constraints (3.25).

**Constraints for traffic congestion:**

$$\sum_{m \in M} F^m_{i_{t_1} j_{t_2}} \leq \left\lfloor C_{i_{t_1} j_{t_2}} \right\rfloor X_{i_{t_1} j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{3.26}$$

$$\sum_{t_2 | \left(i_{t_1}, j_{t_2}\right) \in G} X_{i_{t_1} j_{t_2}} \leq 1, \qquad \forall (i, j) \in L, t_1 \in T \tag{3.27}$$

$$t_1 + \sum_{t \in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \sum_{t \in T} X_{i_{t_2} j_t} (t - t_2) + M \left( 1 - \sum_{t \in T} X_{i_{t_2} j_t} \right), \quad \forall t_1, t_2 \in T, t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij}, (i, j) \in L \tag{3.28}$$

Traffic congestion is expressed through the travel time required to traverse a road link of the network. In the traditional TA problem, travel time is considered a function of traffic flow, and their relationship is described by the BPR function (Dafermos &

Sparrow, 1969): $t = t_0(1 + a(\frac{F}{Q})^b)$ where $F$ is the flow variable, $Q$ denotes the link capacity within an hour, $t_0$ denotes the free-flow travel time, and $a$ and $b$ denotes estimation parameters. However, including this non-linear equation increases the difficulty of solving the MILP model. Thus, we replace the BPR function by imposing several linear constraints which select one from multiple link-traveltime choices at each time point. To realise that, a spatial link capacity $C_{i_{t_1} j_{t_2}}$ that represents the maximum possible flow traversing a certain link $(i, j) \in L$ within a travel time slot between $t_1 \in T$ to $t_2 \in T$ is calculated before the optimisation (Van Essen & Correia, 2019). Firstly, we rewrite the BPR function as $F = Q\left(\frac{1}{a}\left(\frac{t}{t_0} - 1\right)\right)^{\frac{1}{b}}$. Then, the spatial link capacity $C_{i_{t_1} j_{t_2}}$ can be calculated beforehand, and thus can be used as an input parameter, by replacing travel time $t$ by $t_2 - t_1$, $Q$ by $(t_2 - t_1)Q_{ij}$, and $t_0$ by $t_{ij}^{\min}$, which is

$$C_{i_{t_1} j_{t_2}} = (t_2 - t_1)Q_{ij}\left(\frac{1}{a}\left(\frac{t_2 - t_1}{t_{ij}^{\min}} - 1\right)\right)^{\frac{1}{b}}. \tag{3.29}$$

When $t_2 - t_1$ equals the minimum travel time, we add 0.5 to $t_2$ to ensure that the value of $C_{i_{t_1} j_{t_2}}$ is not zero. The spatial link capacity is calculated in advance, providing multiple choices of the link travel time and the corresponding link capacity to the model. Only one link travel time and the corresponding capacity can be selected, as specified by constraints (3.26) and (3.27). Constraints (3.26) impose an additional requirement that the total flow on road link $(i, j)$ never exceeds its spatial link capacity. Constraints (3.28) describe the first-in-first-out (FIFO) rule meaning that the vehicle entering the road link first will leave the road link first. These constraints only apply to time instant $t_1$ and $t_2$ when $t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min}$. Otherwise, if $t_2 > t_1 + t_{ij}^{\max} - t_{ij}^{\min}$, rewritten as $t_2 + t_{ij}^{\min} > t_1 + t_{ij}^{\max}$, it indicates that the arrival time of vehicles entering the road link $(i, j)$ first at time instant $t_1$ with the longest travel time is even earlier than that of vehicles entering the road link $(i, j)$ at a later time instant $t_2$ with the shortest travel time. In this case, there is no need to impose FIFO rule.

## 3.4 Solution methods

In this section, we first propose a two-stage solution method to solve the LLM in Section 3.4.1. Then, in Section 3.4.2, based on the analysis of the relationship between the main decision variables, we adopt a metaheuristic, Parallel Genetic Algorithm (PGA), to obtain a near-optimal solution to the bi-level problem. This method includes an iterative process of solving the lower-level and the upper-level problems.

### 3.4.1 Solution method for the LLM

One question remains to be tackled before we can solve the proposed LLM in Section 3.3.3: how to generate the set of paths $\Pi^r$ for each group of trips $r \in R^{PV}$. The set of paths $\Pi^r$ is referred to as a path pool in the following. After getting the path pool, the proposed LLM can be solved.

Generating all possible paths for a given O-D pair is a hard problem, as its number could be huge, especially in a large-scale network. Solving the proposed model with a large number of alternative paths is not only computationally expensive but also unnecessary. Theoretically, vehicles can drive freely and use any path possible to reach their destination. However, in practice, PVs that drive according to the UE principle will behave selfishly to minimise their travel costs. With this aim, path choices may be limited, as vehicles will always compete for the shortest paths until the shortest one becomes congested and is no longer the optimal one. Then, detouring from the shortest path is needed to avoid traffic congestion and alternative paths will be used. Regarding travel time and travel distance-related costs, long detours are also less likely to occur, which further restricts the available options.

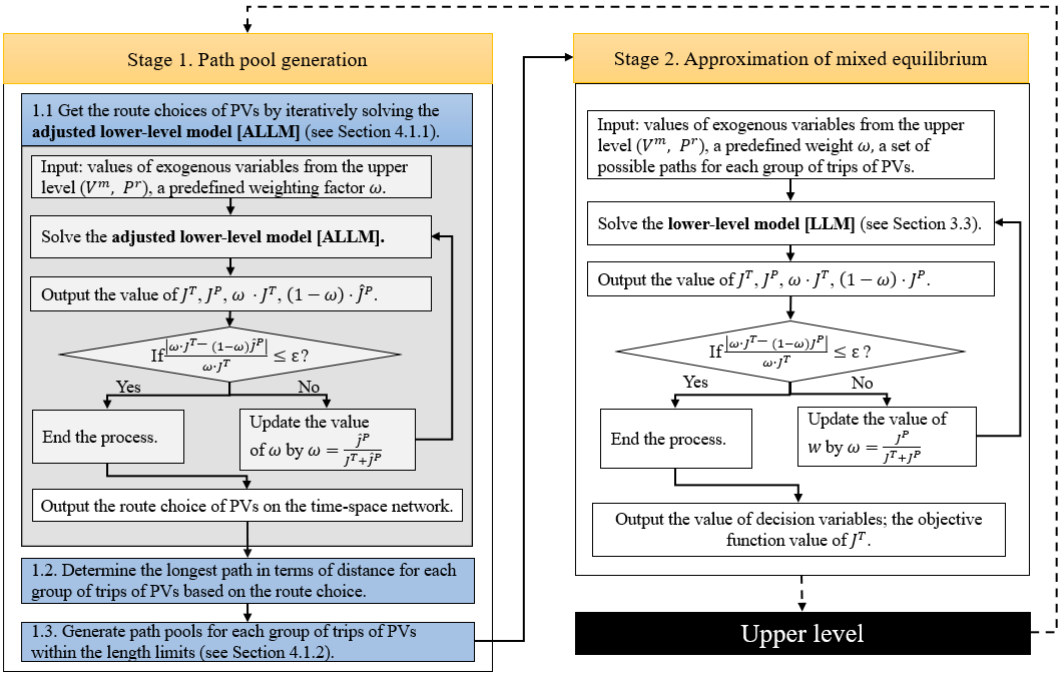



*Figure 3.2: Framework of the lower-level solution method.*

To solve the LLM, we propose the two-stage solution method depicted in Figure 3.2. At Stage 1, we propose a method for generating a path pool for each O-D pair with a reasonable size. The key idea is to identify the longest feasible path in terms of distance that PVs might potentially use, and then generate paths whose length falls beneath the length limit. The longest path for each O-D pair is identified via iteratively solving an additional MILP model which is adjusted from the proposed LLM in Section 3.3.3. The mathematical formulation of this model is presented in Section 3.4.1. The procedure is embedded with the weight determination algorithm described in Appendix 3.A.1. The path enumeration with length limits is presented in Section 3.4.1. By doing so, the unnecessarily long and redundant paths which are unlikely to be used will be eliminated.

At Stage 2, given the path pool for each group of trips, the proposed LLM is solved using the same iterative procedure embedded with the weight determination algorithm. When the algorithm terminates, it is possible to obtain the values of the decision variables and the objective function. These values will be passed to the upper level.

**Adjusted lower-level model (ALLM)**

A new MILP is adjusted from the proposed LLM to produce the longest possible path in terms of distance for PVs in each group of trips. Different from the LLM, the adjusted lower-level model (ALLM) assumes that PVs make route choices based solely on travel times instead of the generalised costs, representing an extreme case where travellers minimise travel time without considering travel distance. While this scenario may not directly correspond to actual travel patterns, the ALLM serves as a crucial step in our solution method to facilitate the solution of the LLM.

The objective function of PVs in ALLM is to minimise the difference between the travel times using different routes for the same O-D pair. By doing so, PVs are likely to detour longer to avoid congestion when a network is super crowded. Later on, when the distance-related cost is included in the objective function of LLM, travellers in PVs will not use paths that are longer than the solution found in the ALLM. Taxis make route choices with the same objective as in LLM. The proposition of ALLM serves as a part of our solution method to facilitate the resolution of the LLM, although PVs only considering travel time may not directly correspond to actual travel behaviours.

Changing the behaviours requires modifying the modelling. As we do not need to track the travel distance using different paths, the path-based variables are no longer necessary in the ALLM. The notations of the newly introduced arc-based variables are presented in Table 3.3. Following is the formulation of the ALLM.

*Table 3.3: Notation*

| Variables | Description |
|---|---|
| $A_t^r$ | Binary variable which is 1 when at least one trip in group $r \in R^{PV}$ arrives at time $t \in T$, and 0 otherwise. |
| $m^r$ | Continuous variable representing the maximum travel time of trips in group $r \in R^{PV}$. |

**Objective function**

$$\textbf{[ALLM]} \qquad \min J = \omega \cdot J^T + (1-\omega) \cdot \hat{J}^P \tag{3.30}$$

where

$$\hat{J}^P = \lambda \sum_{r \in R^{PV}} \frac{m^r}{st^r} + \sum_{r \in R^{PV}} \left( \sum_{\left(i_{t_1}, d_t^r\right) \in G^{PV}} PF^r_{i_{t_1}, d_t^r} \cdot t - a^r \cdot n^r \right) \tag{3.31}$$

The objective function is updated to Equation (3.30), with $J^T$ remaining unchanged from Equation (3.6) and $\hat{J}^P$ represented by Equation (3.31). The aim of routing PVs is to minimise firstly the maximum travel time relative to the shortest possible travel time for all groups of trips and then the total travel time over all the trips. The objective function (3.30) is subject to Constraints (3.8)-(3.17), (3.26)-(3.28), and (3.32)-(3.37).

$$\sum_{j_t \mid \left(o^r_{a^r}, j_t\right) \in G^{PV}} PF^r_{o^r_{a^r} j_t} = n^r, \qquad \forall r \in R^{PV} \tag{3.32}$$

$$\sum_{\left(i_{t_1}, d^r_{t_2}\right) \in G^{PV}} PF^r_{i_{t_1} d^r_{t_2}} = n^r, \qquad \forall r \in R^{PV} \tag{3.33}$$

$$\sum_{j_{t_0} \mid \left(j_{t_0}, i_{t_1}\right) \in G^{PV}} PF^r_{j_{t_0} i_{t_1}} = \sum_{j_{t_2} \mid \left(i_{t_1}, j_{t_2}\right) \in G^{PV}} PF^r_{i_{t_1} j_{t_2}}, \qquad \forall r \in R^{PV}, t_1 \in T,$$
$$t_0 < t_1 < t_2, i \in N^{PV}, i \neq o^r, i \neq d^r \tag{3.34}$$

$$A^r_t \geq \frac{\sum_{i_{t_1} \mid \left(i_{t_1}, d^r_t\right) \in G^{PV}} PF^r_{i_{t_1}, d^r_t}}{n^r}, \qquad \forall r \in R^{PV}, t \in T, a^r \leq t \leq b^r \tag{3.35}$$

$$m^r \geq A^r_t \cdot t - a^r, \qquad \forall r \in R^{PV}, t \in T \tag{3.36}$$

$$F^{PV}_{i_{t_1} j_{t_2}} = \sum_{r \in R^{PV}} PF^r_{i_{t_1} j_{t_2}}, \qquad \forall \left(i_{t_1}, j_{t_2}\right) \in G^{PV} \tag{3.37}$$

Constraints (3.32) and (3.33) ensure that the passenger flows in group of trips $r \in R^{PV}$ depart from the origin node $o^r$ at the scheduled departure time $a^r$ and arrive at the

destination node $d^r$ at time $t \in T$. The flow conservation of passengers driving their PVs is guaranteed by constraints (3.34). The arrival times of trips in group $r \in R^{PV}$ are specified in constraints (3.35) using a binary variable $A_t^r$. Among them, we determine the maximum travel time over the trips in group $r \in R^{PV}$, as indicated in constraints (3.36). The movement of PVs is identical to the movement of travellers within the cars. Constraints (3.37) determine the total vehicle flow on each link in the time-space network.

After solving the ALLM to optimality, the route choices of PVs can be retrieved from the optimal solution, based on which the longest feasible paths in terms of distance for each O-D pair can be identified.

### Path enumeration with length limits

Given the length limitations, the path enumeration method is needed to generate all the paths with lengths shorter than or equal to these limitations. One frequently used path enumeration method is the $k$-shortest path algorithm. Assuming that travellers driving PVs will have perfect information on traffic, going back to a previously visited node is unrealistic. Thus, we adopt a loopless $k$-shortest path algorithm (Yen, 1970) with a predefined sufficiently large value of $k$ ($k$ represents the number of shortest paths to find). The algorithm terminates once the length of a newly generated path exceeds the longest distance threshold. Otherwise, if the total number of generated paths reaches $k$ and the length of the longest path currently found is less than the threshold, we increase the value of $k$ until all paths with lengths less than or equal to the maximum length limits are found.

Using the $k$-shortest path algorithm with a length limit determined by solving model ALLM can effectively restrict the size of the path pool. However, there may be an exception in a particular circumstance. Assuming that vehicles could travel at the maximum permitted speed on the road network without experiencing any congestion, a longer path in terms of distance with a higher maximum speed limit may result in a shorter travel time. It typically occurs outside of built-up areas or on expressways. With a longer length as the threshold value, the $k$-shortest path algorithm is likely to produce a large path pool containing paths that are very similar to one another. Some are deviations from the shortest path, consequently, they are highly overlapped and only differ by a small number of links. These paths are likely to be perceived as the same paths from the driver's perspective as they provide no additional utility. A variety of methods have been proposed for generating a path set considering the overlapping issues. Interested readers can refer to papers written by Chen et al. (2012) and Chondrogiannis et al. (2020).

To shrink the size of the path pool while preserving its heterogeneity, we employ a similarity-based reduction method (Liu et al., 2017; Chondrogiannis et al., 2020). This method consists of removing paths whose similarity to any selected paths exceeds a predetermined threshold $\theta$. Schnabel & Löhse (1997) proposed that the paths are not considered separate if they overlap more than 50%. In this chapter, we use a less restrictive value of 80% to guarantee the solution quality. The similarity between two paths is calculated by dividing the total length of overlapping links by the length of the shorter path between them. In this way, the unnecessarily lengthy paths could be excluded. The pseudo-code of the similarity-based path pool reduction procedure can be found in Algorithm 3.1. By reducing the number of possible paths in the path pools, the number of variables and constraints in the LLM are reduced.

**Algorithm 3.1** Similarity-based path pool reduction procedure

**Require:** similarity threshold $\theta$, longest distance thresholds $ld^r$ for $r \in R^{PV}$.
**Ensure:** $PathPoolUpdated$ (a list).
Initialise empty lists $PathPool := [[]]$ for $r \in R^{PV}$, $PathPoolUpdated := [[]]$ for $r \in R^{PV}$.
**for** $r$ **in** $R^{PV}$ **do**
    Generate paths within the longest distance thresholds $ld^r$ and sort them by path length from shortest to longest.
    Save the sorted paths to list $PathPool[r]$.
    Add the shortest path to list $PathPoolUpdated[r]$.
    **for** $path1$ **in** $PathPool[r]$ **do**
        flag := **true**
        **for** $path2$ **in** $PathPoolUpdated[r]$ **do**
            Compute the similarity $\theta'$ between $path1$ and $path2$.
            **if** $\theta' > \theta$ **then**
                flag := **false**
                **break**
            **end if**
        **end for**
        **if** flag **is true then**
            Add $path1$ to list $PathPoolUpdated[r]$.
        **end if**
    **end for**
**end for**

## 3.4.2 Solution method for the overall problem

To solve the proposed bi-level programming model, an overall algorithm is required after solving the lower-level model. In our problem, the upper level is relatively straightforward compared to the lower level due to the limited number of decision variables

(fleet size variables for CTs and ATs) and constraints. While a simple enumeration scheme-based method, such as a binary search algorithm, appears to be a possibility, this is not suitable for solving a heterogeneous FSMVRP considering endogenous traffic congestion and the interaction of different types of vehicles. We explain the reasons below.

First, the interdependence of the fleet size variables increases complexity. Modifying one variable can potentially lead to changes in the other variable since the fleet sizes directly impact road traffic and congestion. Additionally, this relationship is nonlinear and non-monotonic, which means that multiple local minima may exist. For instance, one local minimum could occur when both fleet sizes are small, while another local minimum could be found when the AT fleet size is large, and the CT fleet size is even smaller. In the latter case, with more ATs, relocation needs can be reduced, thus alleviating congestion effects on the road network. Consequently, a smaller fleet of CTs would suffice to serve more requests, leading to cost savings for TNCs as they employ fewer drivers for CTs. A binary search algorithm cannot be used in our case, as it discards half of the feasible region once the searching direction is determined. Consequently, it may only find one local minimum while another local minimum may exist in the discarded feasible region. Therefore, relying on a binary search algorithm to find all possible local minima is not possible.

Enumerating all feasible solutions is a possible, but computationally expensive approach, particularly when the fleet size bounds are large and there are multiple types of fleets. Given these considerations, employing heuristic/meta-heuristic methods to solve the proposed bi-level problem is more suitable. These methods can effectively handle the complexities of the problem and are better equipped to identify multiple local minima, considering the nonlinearity and non-monotonicity of the relationship between fleet sizes and congestion. Several heuristic and meta-heuristic techniques have been employed to address bi-level leader-follower problems, such as genetic algorithm (Madadi et al., 2020), simulated annealing (Chen et al., 2017), tabu search (Camacho-Vallejo et al., 2021), etc. Among them, the genetic algorithm is one of the most commonly used methods (Farahani et al., 2013) and has been shown to have a competitive performance compared with other methods (Liu et al., 2009).

The primal disadvantage of adopting a Genetic Algorithm (GA) in our problem is the computationally expensive fitness evaluation process for each individual in the population along with the evolution process. However, since GA is a population-based meta-heuristic working on improving the quality of the whole population instead of a single solution, every individual can be evaluated independently at each generation. The independent parts of GA can be distributed to different processes and executed in parallel to reduce computational time. Interested readers may refer to the literature for

more details (Eklund, 2004; Katoch et al., 2021). In this chapter, we adopt a method called Global single-population master-slave GA which parallelises the fitness evaluation process (solving the lower-level problem) because it is the most time-consuming part of the problem.

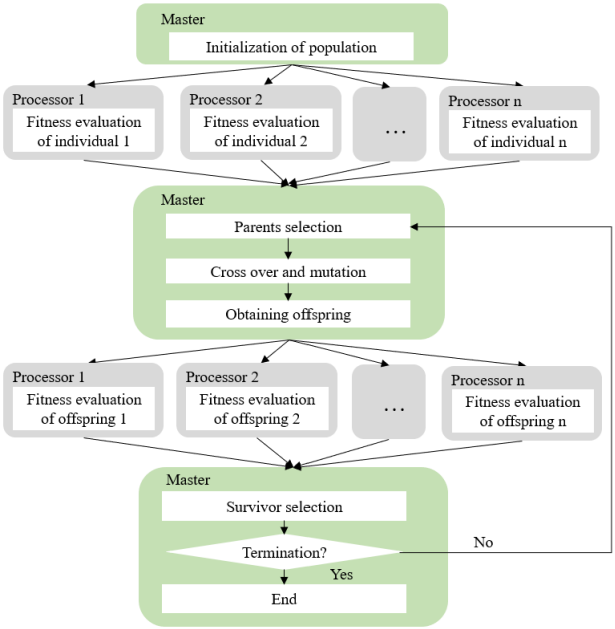



*Figure 3.3: Structure of the parallel genetic algorithm (PGA)*

GA is firstly applied at the upper level to generate individuals, which are then distributed to independent processors to solve the lower-level problem. No tasks associated with the GA process such as crossover and mutation operators are paralleled as its execution takes a very short time. Parallelism enables the use of a multi-core CPU's computational capacity, resulting in a significant reduction in computational time. Figure 3.3 shows the structure of the parallel genetic algorithm (PGA). A brief overview of the PGA is presented in the following section.

**Initialisation** The first step of the PGA is to initialise the population. The population consists of a certain number of chromosomes, each of whom represents a potential solution to our problem. In this chapter, we simplify the problem by assuming that no trips will be rejected. Thus, each chromosome is composed of two integer variables $[V^{CT}, V^{AT}]$, representing the fleet size of CTs and ATs.

Before randomly generating the population's first generation, the bounds for these

two variables need to be specified. One upper bound for the fleet size of CTs and ATs is the total number of trips for CTs and ATs, which means one vehicle per trip, while a lower bound is not that easy to find. We search for respective lower bounds of CTs and ATs that ensure the feasibility of the model. In other words, these values are the minimum number of vehicles below which it would not be possible to satisfy the demand. Thus, these lower bounds correspond to the minimum fleet sizes for the problem. A binary search algorithm is proposed to find that lower bound. Notice that the binary search will be conducted on only one type of fleet at a time, with the value of the other type being its upper bound, to make sure that the latter type never introduces infeasibility. Given the fleet size value of CTs and ATs, the feasibility of the model can be identified by solving the LLM (not necessarily to the optimum). This feasibility can then act as the indicator to repeatedly divide the fleet size bound of CTs or ATs that contain the minimum feasible solution in half until there is only one value remaining. This value is the lower bound of one fleet.

An initial lower bound needs to be given before implementing the binary search algorithm. We assume all the passengers will be delivered in the shortest possible travel time and no relocation time of taxis is considered. Once the passenger is dropped off at the destination, the taxi can immediately begin serving the next trip. Thus, this initial lower bound value can be obtained by finding the maximum number of overlapping travel time intervals for all trips at any point in time. Here, the travel time interval for each trip is defined as the time difference between the departure time and the earliest possible arrival time when heading from the origin to the destination. Figure 3.4 illustrates how to determine the minimum number of taxis required to serve four trips. In this case, the maximum number of overlapped travel time intervals is three, implying that three vehicles are needed as a minimum to serve all trips. The pseudo-code of the detailed process for finding the lower bound of fleet sizes can be found in Appendix 3.B. Knowing the bound of the fleet size of CTs and ATs, the population in the first generation can be randomly generated from a uniform distribution.

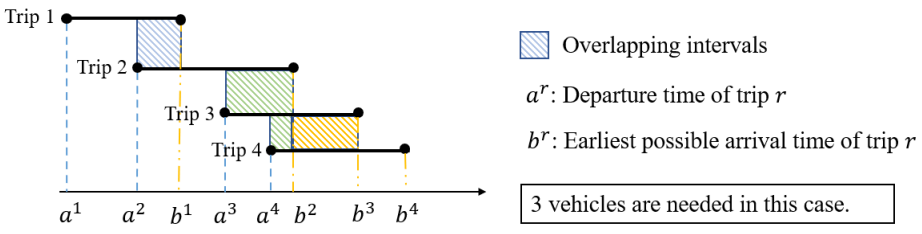



*Figure 3.4: Illustration of finding the minimum number of taxis to serve four trips.*

**Parents selection** The parents who will have offspring are selected from the population using a fitness proportionate selection method. Knowing the fitness value of each individual, we rank the individuals and then introduce a new fitness function based on the rank. Individuals with a higher rank are more likely to be selected as parents.

**Crossover operator** The crossover operator exchanges the chromosomes of the selected parents to produce two offspring. In our case, the crossover operator is applied with a probability $P_c$. We randomly generate a number between zero and one for each pair of parents to determine whether we should apply this operator. If this random number is less than $P_c$, we perform the crossover operator. Otherwise, we keep the parents' chromosomes unchanged. In this chapter, we cross the fleet size values to change the chromosome of the parents as only two values are included in each chromosome.

**Mutation operator** After the crossover operator is applied, the mutation operator is executed for every offspring with a given probability. Two types of mutation operators are used in our algorithm: the creep mutation operator and the random mutation operator. In our case, a simplified creep mutation operator is used by simply performing +1 or -1 to each value in a chromosome with an equal probability. By doing so, the algorithm could exploit more solutions in a concentrated area in the solution space. The random mutation operator is used to explore a large region for a better solution and avoid the local optima. It replaces the value in the chromosome with a random integer between the upper bound and lower bound of the fleet size with a given probability.

The mutation operator is applied to fleet size values from each chromosome randomly. For the newly produced offspring, we perform the creep mutation operator. If the chromosome of an offspring has already existed in the current population, the creep mutation operator is applied with a high probability $P_{cm1}$. Else, the creep mutation operator is applied with a low probability $P_{cm2}$. For the parents whose chromosomes stay unchanged after performing the crossover, the random mutation operator is applied with a probability $P_{rm}$ to explore the feasible region. After performing the mutation operator, the chromosomes will be added to the list of offspring if no individual in the current population has the same chromosomes as them.

**Fitness evaluation** Once we obtain new offspring, a fitness evaluation will be conducted. To avoid performing repetitive calculations, the check is made to see if the fitness of the current offspring has been calculated previously. For those who have been computed, we can obtain their fitness value directly from memory. For those offspring who have never been evaluated, individual fitness evaluations will be distributed

to different processors and performed in parallel to maximise the computational capacity of multiple cores.

Multiple criteria are defined to terminate the LLM and ALLM solution process in case the computational time is extremely long. Firstly, the model is solved as close to optimality as possible within a small time limit (denoted as a soft time limit). After reaching this time limit, the model is terminated either because the MIP gap reaches a predefined gap limit or the computational time reaches a predefined large time limit (denoted as a hard time limit).

**Survivor selection** The elitism replacement approach is used for the survivor selection. After getting the fitness value of the offspring, the previous generation and the offspring are put in a pool. The first $q\%$ best individuals in terms of fitness value are firstly selected. Then, we randomly select from the rest individuals until the number of selected individuals equals the predefined population size.

**Termination criteria** We terminate the algorithm based on three criteria. First, if there is no improvement of the best individual in the population for a certain number of successive iterations. Second, if the average population quality of the top 5 fittest individuals has no improvement after a certain number of successive iterations. Here, we measure the average population quality using the mean and standard deviation values of the individual fitness. Third, if the predefined maximum number of generations has been reached.

## 3.5 Computational experiments

To test the performance of the proposed model and algorithm, we present two case studies in this section. Firstly, a small toy network case study is presented to demonstrate that solving the proposed lower-level problem can achieve an approximated mixed-equilibrium in Section 3.5.1. Then, in Section 3.5.2, we apply the proposed bi-level model to a quasi-real case study representing the city of Delft, in the Netherlands.

### 3.5.1 Demonstration of the lower-level problem on a small toy network

The small toy network we use contains 16 nodes and 48 directed links (each road segment has two directions), as shown in Figure 3.5. Among all nodes, nodes 4, 6, 9

and 11 are parking nodes that can be regarded as free parking depots, while the rest of the nodes do not allow parking. For the links in this small toy network, each of them has an equal length of 2 kilometres and the same capacity of 1800 vehicles/h. The minimum and the maximum travel time for traversing a link are set to 1 time step (2.5 minutes) and 4 time steps (10 minutes), respectively. In the current experiment, no AVs-only zone is included as we would like to leave out the impact of the AVs zone on the route choices and only focus on the equilibrium achieved by solving the model.

Six groups of trips are considered, with the trip information shown in Table 3.4. Here, for simplicity, only CTs and PVs are considered options for travellers because the routing behaviours of CTs and ATs are the same. By doing so, we focus on comparing the route choices of road users with different routing behaviours (SO and UE). The lower bound for the CTs' fleet size can be easily derived from the given data, 390, as all the trips depart at the same time. Given a great number of trips in each group, traffic congestion occurs in the network.

The parameters related to the CTs and PVs are as follows. $co^m$ with $m \in \{CT, PV\}$ is set to 0.25 euros/km and 0.27 euros/km, respectively, representing the unit operational costs for using CTs and PVs. These values are calculated according to the methodology proposed by Bösch et al. (2018). $cd$ represents the drop-off delay penalty, which is 0.2 euros/min based on Liang et al. (2020). $ct$ is the travel time related cost for PVs which is set to 9 euros/h based on Kouwenhoven et al. (2014). The estimation parameters $a$ and $b$ of the BPR function are set to 2 and 4, respectively, based on Van Essen & Correia (2019). The optimisation period is 10 time instants.

Using the minimum fleet size of CTs as the input, the LLM is solved to demonstrate the approximated mixed equilibrium. A base scenario (S0) is tested first, followed by two different scenarios to see how the value of the delay penalty affects the route choice of different road users when reaching an approximated mixed equilibrium. In the first scenario (S1), we assume there is no penalty for delivery delay, so $cd$ is set to 0 euros/min. In the second scenario (S2), a high penalty for delivery delay is set to 0.4 euros/min. Here, only the parameters that could be controlled by TNCs are tested. The operational costs of CTs and PVs, and the value of travel time using PVs are not varied for sensitivity analysis as these parameters could be well estimated (Kouwenhoven et al., 2014; Bösch et al., 2018).

The lower-level framework was implemented in Python and solved using Gurobi 9.0.2 on an Intel(R) Xeon(R) W-2123 CPU @3.60 GHz, and 32.00 GB RAM computer. The base scenario was tested firstly with a given initial weight $\omega$ as 0.5. The algorithm terminates when the relative difference between the contributed values is smaller than 5%.

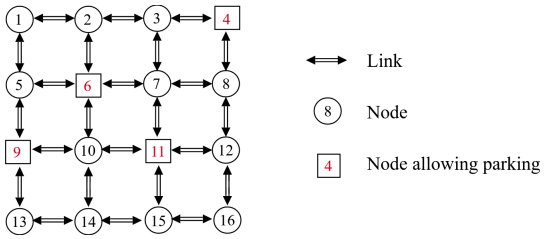



*Figure 3.5: Illustration of the small toy network.*

*Table 3.4: Demand for CTs and PVs.*

| Index of group of trips | Origin | Destination | Departure time (time instant) | Latest arrival time (time instant) | Number of trips | Type of vehicle |
|---|---|---|---|---|---|---|
| 1 | 1 | 10 | 0 | 10 | 140 | PV |
| 2 | 1 | 10 | 0 | 10 | 140 | CT |
| 3 | 11 | 6 | 0 | 10 | 130 | PV |
| 4 | 11 | 6 | 0 | 10 | 130 | CT |
| 5 | 5 | 7 | 0 | 10 | 120 | PV |
| 6 | 5 | 7 | 0 | 10 | 120 | CT |

### Computational results at Stage 1: Path pool generation

The computational results are shown in Table 3.5, demonstrating that three iterations are needed to satisfy the convergence criterion and accurately determine the value of $\omega$ in stage 1. After solving the ALLM, we retrieve the route choices of CTs and PVs from the optimal solution and then display the results in Table 3.6. From the table, we observe that PVs choose different paths with the same travel times. An equilibrium state is reached in which no driver is able to deviate from his/her current route otherwise travel time will increase. Hence, this scenario exemplifies a UE. In the case of the taxis, the travel times and distances differ from each other. Some taxis take the shortest path regarding length and travel time, while others are sacrificed to reach a SO. Compared with the PVs, taxis would prefer shorter paths in terms of distance as they consider generalised costs when routing. But PVs choose longer paths to have shorter travel times.

The longest travel distance of PVs for group 1, 3 and 5 can be determined from the optimal solution of ALLM, which are 10km, 8km, and 8km, respectively. These values are then used as the length limits to generate a path pool for each group of trips using the k-shortest path algorithm. In this small case, 9 paths, 6 paths and 7 paths are obtained for group 1, 3 and 5, respectively, which are used for Stage 2.

*Table 3.5: Results of the base scenario in the small toy network.*

| | Number of iterations | Value of $\omega$, $\lambda$ | Objective function values | Contributed values |
|---|---|---|---|---|
| Stage 1 | 3 | 0.99987<br>1251600 | $J^T$ = **1030**,<br>$\hat{J}^P$ = 7928780 | $\omega \cdot J^T$ = 1029.87,<br>$(1-\omega) \cdot \hat{J}^P$ = 1029.87 |

*Table 3.6: Route choices at Stage 1.*

| O-D | Model | Paths | Flow | Path length (km) | Travel time (timestep) |
|---|---|---|---|---|---|
| 1-10 | ALLM | Taxi (SO): [1-2-6-10], [1-5-6-10], [1-5-9-10]<br>PV (UE): [1-2-3-7-11-10], [1-2-6-10] | 34, 53, 53<br>48, 92 | 6, 6, 6<br>10, 6 | 7, 4, 4<br>7, 7 |
| 11-6 | ALLM | Taxi (SO): [11-10-6], [11-7-6]<br>PV (UE): [11-12-8-7-6], [11-15-14-10-6], [11-7-6] | 53, 77<br>53, 53, 24 | 4, 4<br>8, 8, 4 | 2, 4<br>4, 4, 4 |
| 5-7 | ALLM | Taxi (SO): [5-6-7], [5-1-2-6-7], [5-1-2-3-7]<br>PV (UE): [5-6-7], [5-9-10-11-7] | 59, 8, 53<br>67, 53 | 4, 8, 8<br>4, 8 | 4, 6, 6<br>4, 4 |

**Computational results at Stage 2: Approximation of mixed equilibrium**

Knowing the path pool for each group of trips, the LLM is solved. The final results, displayed in Table 3.7, reveal that three iterations are required to achieve a balanced contribution of the objective function between taxis and PVs, signifying the convergence of the algorithm. From the results, we see that the total operational cost of taxis in the LLM, denoted by $J^T$ is higher than that in the ALLM, because of the greater travel time and longer travel distance of CTs resulting from the intense competition for the lowest cost paths with PVs.

*Table 3.7: Final results of the base scenario in the small toy network.*

| Stage 2 | Number of iterations | Value of $\omega$, $\lambda$ | Objective function values | Contributed values |
|---|---|---|---|---|
| LLM | 3 | 0.99953<br>591322.41 | $J^T$ = **1192**,<br>$J^P$ = 2583009.41 | $\omega \cdot J^T$ = 1191.44<br>$(1-\omega) \cdot J^P$ = 1207.93 |

Table 3.8 shows the final route choices of taxis and PVs. In the LLM, PVs consider the general cost when making route choices. From the table, we can see that PVs

*Table 3.8: Final route choices.*

| O-D | Model | Paths | Flow | Path length (km) | Travel time (timestep) |
|---|---|---|---|---|---|
| 1-10 | LLM | Taxi (SO): [1-2-6-10], [1-5-9-10]<br>PV (UE): [1-2-6-10], [1-5-6-10] | 87, 53<br>39, 101 | 6, 6<br>6, 6 | 6, 5<br>6, 7 |
| 11-6 | LLM | Taxi (SO): [11-12-8-7-6], [11-7-6]<br>PV (UE): [11-7-6], [11-10-6] | 8, 122<br>4, 126 | 8, 4<br>4, 4 | 4, 4<br>4, 4 |
| 5-7 | LLM | Taxi (SO): [5-1-2-3-7], [5-6-7], [5-9-13-14-10-6-7], [5-9-10-6-7]<br>PV (UE): [5-6-7] | 53, 6, 8, 53<br>120 | 8, 4, 12, 8<br>4 | 4, 4, 7, 5<br>4 |

choose paths with similar or the same generalised costs. Taxis take paths with diverse generalised costs. Some taxis are sacrificed and take a path with a large cost to reach a SO. By analysing the flow patterns and the route choices of CTs and PVs, we can demonstrate that an approximated mixed equilibrium has been reached.

#### Sensitivity analysis

A sensitivity analysis regarding the delay penalty parameter *cd* is carried out. For illustration purposes, only the route choices of CTs and PVs departing from node 11 and heading to node 6 are shown in Table 3.9. Similar patterns happen for the other O-D pairs. When there is no delay penalty in scenario 1, taxis no longer care about the travel time and only consider the travel distance. Therefore, in the ALLM, taxis choose the shortest distance path with a long travel time, while in the LLM, PVs would also like to join in the competition for the shortest travel distance. To cope with the needs of PVs, the travel time of the shortest paths can no longer be very long. Consequently, some taxis have to divert to longer paths to avoid extreme congestion. In scenario 2, where the delay penalty is twice as high, we found that there is no change to the route choices of PVs and taxis in the ALLM, while in the LLM, taxis prefer to use longer paths but lower travel time to reduce the delay penalty.

### 3.5.2 Quasi-real case study of the city of Delft, in the Netherlands

#### Application setting

The next set of experiments is based on the network of the city of Delft, which is located in the South Holland province of the Netherlands. We call this case study a

*Table 3.9: Computational results for the referred scenarios*

| Scenario | Model | Paths | Flow | Path length (km) | Travel time (timestep) |
|---|---|---|---|---|---|
| S0 (Base) | ALLM | Taxi (SO): [11-10-6], [11-7-6] | 53, 77 | 4, 4 | 2, 4 |
| | | PV (UE): [11-12-8-7-6], [11-15-14-10-6], [11-7-6] | 53, 53, 24 | 8, 8, 4 | 4, 4, 4 |
| | LLM | Taxi (SO): [11-7-6], [11-12-8-7-6] | 122, 8 | 4, 8 | 4, 4 |
| | | PV (UE): [11-7-6], [11-10-6] | 4, 126 | 4, 4 | 4, 4 |
| S1 (No delay penalty) | ALLM | Taxi (SO): [11-10-6] | 130 | 4 | 7 |
| | | PV (UE): [11-12-8-7-6], [11-7-6], [11-15-14-10-6] | 24, 53, 53 | 8, 4, 8 | 4, 2, 4 |
| | LLM | Taxi (SO): [11-10-6], [11-12-8-7-6] | 122, 8 | 4, 8 | 4, 8 |
| | | PV (UE): [11-7-6], [11-10-6] | 126, 4 | 4, 4 | 4, 4 |
| S2 (High delay penalty) | ALLM | Taxi (SO): [11-10-6], [11-7-6] | 77, 53 | 4, 4 | 4, 2 |
| | | PV (UE): [11-12-8-7-6], [11-15-14-10-6], [11-10-6] | 53, 53, 24 | 8, 8, 4 | 4, 4, 4 |
| | LLM | Taxi (SO): [11-12-8-7-6], [11-7-6], [11-15-14-10-6] | 53, 49, 28 | 8, 4, 8 | 4, 2, 4 |
| | | PV (UE): [11-10-6], [11-7-6] | 126, 4 | 4, 4 | 4, 2 |

quasi-real one, because of the following reasons: (1) A simplified road network of Delft is used instead of the real one; (2) The expansion process and the transformed links of the AVs-only zone are experimental; (3) Despite using as source real travel data, the mobility data tested in the case study was generated from the Dutch mobility dataset (MON 2007/2008) which does not have a large sample for this city (Correia & Van Arem, 2016). The purpose of carrying out this case study is to test the effectiveness of the proposed method and get first insights into the impacts on travellers imposed by AVs-only zones.

The road network used for this study is simplified to 35 nodes and 104 directed links (each road segment has two directions). In the network, nodes 19, 3, 10, 22, 27 and 15 are designated as free parking depots for taxis. Both the CTs and ATs are permitted to utilise the nodes located at the border of the AVs-only zone. Moreover, two types of links with one or two lanes per direction and a capacity of 1600 or 3200 are considered. The maximum travel speed for the lower and higher capacity links was assumed to be 50km/h and 70km/h, respectively. The road capacity triples after the road links are transformed to AV links. The minimum travel time and maximum travel time on each link are calculated based on the free-flow speed and a speed of 5 km/h.

Figure 3.6 (a) depicts the conventional road network where there is no AVs-only zone. The AVs-only zone is expanded gradually covering 25%, 50%, 75%, and 100% of the links, as shown in Figure 3.6. To expand the AVs-only zone, we initially define it in areas characterised by frequent traffic congestion, such as the city centre, train


*Figure 3.6: Road networks of Delft with different AVs-only zone size: (a) 0%, (b) 25%, (c) 50%, (d) 75%, (e) 100%.*

station, and university campus. Subsequently, we employ a randomised approach to gradually expand the zone until it encompasses the entire city. However, it is important to note that the optimal design of the AVs-only zone is beyond the scope of this thesis. At that point, no HVs are permitted to operate on the network. For this particular exceptional scenario, the fleet sizing problem can be easily solved by a single-level MILP model with the objective function (3.1) subject to Constraints (3.3), (3.4), (3.8)-(3.17) and (3.26)-(3.28).

The Dutch mobility dataset (MON 2007/2008) is used in this study to generate mobility data for the morning peak hour. This data includes trip information, such as

origin, destination, departure time, arrival time, and travel mode for O-D pairs on a typical working day. A total of one hour is studied during the morning peak when demand is high and traffic congestion has a significant impact on vehicles' route choices. The data set we used includes 1163 trips in total, with 23 groups for taxis and 23 groups for PVs. The departure time of each group of trips is distributed within one hour. Once generated, the departure time does not change with the expansion of the AVs-only zone. Regarding the preference of CTs and ATs, in a base scenario with 0% AVs-only zone, more than 80% of the trips with a preference for CTs are generated assuming that the trust of users towards AVs in level 5 is relatively low at the early stage (Correia et al., 2019). Besides, a time step of 2.5 mins is used.

The parameter values used in the solution method are shown in Table 3.10. For simplification purposes, the minimum service rate $\alpha$ in Constraint (4) is set to 1 in this case study, meaning that all demand will be served by taxis. The influence of the value of $\alpha$ will be studied in future research. The appendix contains the parameter tuning for the similarity threshold and population size.

*Table 3.10: Parameter values.*

| Parameters | Values | Parameters | Values |
|---|---|---|---|
| $co^m, m \in \{CT, AT, PV\}$ | 0.25, 0.32, 0.27 euros/km | Population size | 8 |
| $cp$ | 10 euros/hour | Crossover rate | 0.8 |
| $cf^m, m \in \{CT, AT\}$ | 1, 1.2 euro/vehicle/h | Mutation rate ($P_{cm1}$, $P_{cm2}$, $P_{rm}$) | 0.5, 0.03, 0.5 |
| $cd$ | 0.2 euros/min | Percentage of elitism individuals | 0.8 |
| $ct$ | 0.27 euros/min | Maximum number of generations | 100 |
| $p^0$ | 3 euros/trip | Maximum number of iterations with no change for the best solution | 20 |
| $p^m, m \in \{CT, AT\}$ | 2.55, 2.3 euros/km | Similarity threshold ($\theta$) | 80% |
| Minimum service rate ($\alpha$) | 1 | Relative difference threshold ($\varepsilon$) | 5% |
| Soft time limit | 30 mins | Maximum number of iterations with no change for the quality of the top five fittest individuals | 10 |
| Hard time limit | 60 mins | MILP gap limit | 2% |

### Performance of the solution method

We applied the proposed solution method to the bi-level problem in several scenarios where the coverage rate of the AVs-only zone is 0%, 25%, 50% and 75%. Figure 3.7 shows the computational performance in each scenario. Three main indicators are shown along with the iteration until the algorithm terminates: the best fitness value, the mean and the standard deviation value of the fitness value of the top five fittest individuals.

According to the charts, convergence has been reached for all four scenarios. In addition, the solution method ended because the maximum number of iterations where the mean value and the standard deviation of the top five fittest individuals do not change has been reached. In the first few iterations, PGA explored the feasible solution space and selected the best few individuals to produce the next generation. As the iterations progressed, the mean fitness of the top five fittest individuals approached the best fitness value, and their standard deviation approached zero. This means that the quality of the population has reached a stable and favourable state in a limited number of iterations. The computational times for these four scenarios are 23.7h, 13.6h, 4h, and 6.7h, respectively, demonstrating a decreasing trend as the coverage of the AVs-only zone increases. Besides, to mitigate the risk of the algorithm converging to a local optimum, we executed the PGA algorithm multiple times using identical experimental settings for each scenario. All yielded consistent results.

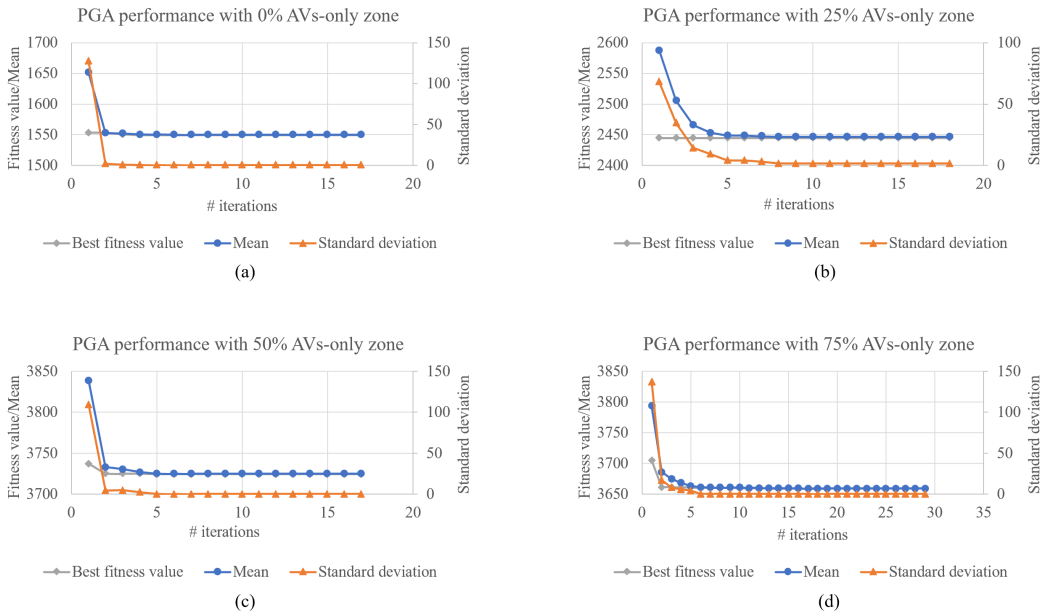



*Figure 3.7: Performance of the solution method with different coverage rates of the AVs-only zone: best fitness value, mean and standard deviation of the fitness value of the top five fittest individuals.*

### Comparison between the best fleet size and the lower bound

The optimisation results for the base scenario are shown in Table 3.11. Note that the fleet size obtained by applying PGA is a near-optimal solution as the optimality cannot be guaranteed since PGA is a meta-heuristic. We call it 'best' hereinafter to distinguish it from its lower bound. The lower bound which is the minimum feasible fleet size to satisfy all the demand in different AVs-only zone settings can be obtained by applying the binary search algorithm presented in Section 3.4.2.

Looking at the fleet size in each scenario, we notice that the minimum fleet size is the best one when the coverage rate of the AVs-only zone is 25%, whereas, in the remaining scenarios, the best fleet size differs from the minimum one. To be more specific, only the best fleet size of the ATs differs. From Table 3.11, it is quite clear that this difference comes from the cost-saving deriving from the shorter relocation distance of ATs, despite the larger fleet size. In all the scenarios, the best fleet size of CTs equals their minimum feasible fleet size, as deploying a larger CT fleet is more costly because more human drivers have to be hired. That is why a TNC will try to deploy the least number of CTs. Therefore, deploying ATs may create a cheaper form of on-demand mobility.

The relocation distance consists of three possible parts: the distance from the drop-off location to the parking depot, the distance from the parking depot to the next pick-up location and the distance from the drop-off location to the next pick-up location, which therefore highly depends on the location of the parking depots and the demand pattern. Theoretically, locating a parking depot in an area frequently visited by travellers or densely populated could reduce the relocation distance. However, such locations typically lack sufficient space for constructing large parking facilities. In this case study, three parking depots are located in densely populated areas (corresponding to nodes 3, 10 and 22), and four parking depots are located on the outskirts or outside the city (corresponding to nodes 11, 15, 19, 27). Less densely distributed parking depots also result in large relocation costs. Nevertheless, the optimal location and distribution of parking depots are not the focus of this chapter.

### Demonstration of the approximated user equilibrium

To demonstrate that the approximated user equilibrium for PVs has been achieved, we calculate the ratios of the maximum cost to the minimum cost among all the utilised paths for each group of trips. A ratio approaching 1 indicates superior results, as it signifies that the costs of all utilised paths are similar. Then, in Figure 3.8, we show the mean and standard deviation (SD) of the calculated cost ratios across all groups of trips for scenarios with different coverage rates of the AVs-only zone (0%, 25%, 50%,

*Table 3.11: Optimisation results with different coverage rate of AVs-only zones ("M" represents the minimum feasible fleet size; "B" represents the best fleet size).*

| Coverage rate | 0% | | 25% | | 50% | | 75% | | 100% | |
|---|---|---|---|---|---|---|---|---|---|---|
| | M | B | M | B | M | B | M | B | M | B |
| Fleet size of CTs | 95 | 95 | 89 | | 27 | 27 | 7 | 7 | 0 | 0 |
| Fleet size of ATs | 32 | 43 | 253 | | 550 | 659 | 608 | 714 | 662 | 711 |
| Total profit (euros) | 2581.52 | 2588.67 | 6591.36 | | 15539.92 | 15601.70 | 16803.63 | 16861.36 | 17643.82 | 17674.55 |
| Total cost (euros) | 1562.62 | 1555.46 | 2445.59 | | 3811.08 | 3749.30 | 3734.22 | 3676.50 | 3682.47 | 3651.74 |
| Number of trips for CTs | 166 | | 106 | | 34 | | 14 | | 0 | |
| Number of trips for ATs | 47 | | 424 | | 920 | | 1072 | | 1163 | |
| Utilisation rate of CTs | 1.75 | 1.75 | 1.19 | | 1.26 | 1.26 | 2 | 2 | - | - |
| Utilisation rate of ATs | 1.47 | 1.09 | 1.68 | | 1.67 | 1.39 | 1.76 | 1.50 | 1.76 | 1.63 |
| Total travel distance of CTs (km) | 1201.41 | 1201.41 | 1380.05 | | 205.39 | 205.39 | 72.28 | 72.28 | - | - |
| Total travel distance of ATs (km) | 516.20 | 452.60 | 2349.88 | | 8300.75 | 7698.94 | 8838.89 | 8261 | 9025.22 | 8745.45 |
| Total travel distance of PVs (km) | 6414.72 | 6414.72 | 7091.97 | | 975.91 | 975.91 | 247.25 | 247.25 | - | - |
| Percentage of deliver distance of CTs (%) | 86.52 | 86.52 | 89.84 | | 74.16 | 74.16 | 62.19 | 62.19 | - | - |
| Percentage of deliver distance of ATs (%) | 76.79 | 87.58 | 82.08 | | 86.32 | 93.06 | 85.37 | 91.35 | 85.93 | 88.68 |
| Percentage of relocate distance of CTs (%) | 13.48 | 13.48 | 10.16 | | 25.84 | 25.84 | 37.81 | 37.81 | - | - |
| Percentage of relocate distance of ATs (%) | 23.21 | 12.42 | 17.92 | | 13.68 | 6.94 | 14.63 | 8.65 | 14.07 | 11.32 |
| Percentage of detour distance of CTs (%) | 1.86 | 1.86 | 33.62 | | 26.99 | 26.99 | 0 | 0 | - | - |
| Percentage of detour distance of ATs (%) | 0 | 0 | 0.29 | | 1.27 | 1.37 | 0.94 | 1 | 0 | 0 |
| Percentage of detour distance of PVs (%) | 1.27 | 1.27 | 28.77 | | 39.22 | 39.22 | 0 | 0 | - | - |
| Total delayed time of CTs (time step) | 38 | 38 | 78 | | 44 | 44 | 0 | 0 | - | - |
| Total delayed time of ATs (time step) | 0 | 0 | 64 | | 252 | 252 | 163 | 163 | 0 | 0 |
| Total delayed time of PVs (time step) | 418 | 418 | 549 | | 0 | 0 | 0 | 0 | - | - |
| Average delayed time per trip of CTs (time step) | 0.23 | 0.23 | 0.74 | | 1.29 | 1.29 | 0 | 0 | - | - |
| Average delayed time per trip of ATs (time step) | 0 | 0 | 0.15 | | 0.27 | 0.27 | 0.15 | 0.15 | 0 | 0 |
| Average delayed time per trip of PVs (time step) | 0.44 | 0.44 | 0.87 | | 0 | 0 | 0 | 0 | - | - |
| MILP gap value (%) | 4.63 | 2.8 | 0 | | 0 | 0 | 0 | 0 | 0 | 0 |

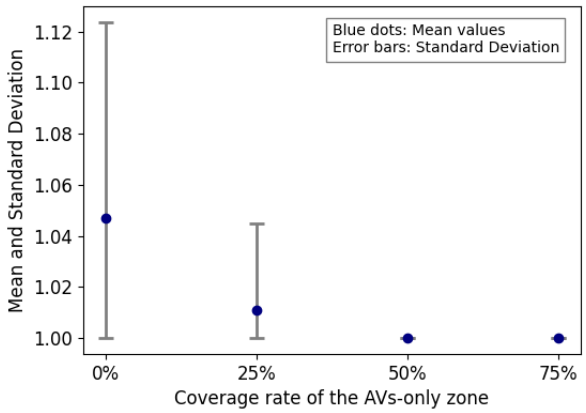



*Figure 3.8: Mean and standard deviation of the cost ratios across all groups of trips for scenarios with different coverage rates of AVs-only zone (0%, 25%, 50%, and 75%) and the best fleet sizes.*

and 75%) and the best fleet sizes.

As illustrated in Figure 3.8, the mean values range between 1 and 1.047 for all the scenarios. This indicates that, on average, the costs of the utilised paths are very similar across each group. For scenarios with a 0% and 25% coverage rate of the AVs-only zone, the SD values are 0.077 and 0.034, respectively, as represented by the error bars in the figure. These values are reasonable, considering a perfect UE can hardly be achieved because of the discrete time setting in the time-space network. Notably, when the AVs-only zone coverage rate exceeds 50%, all scenarios exhibit a mean value of 1 and an SD of 0. This suggests that UE has been achieved without any deviation in the groups. Additionally, the mean and SD values show a decreased trend in the figure with the increased coverage rate of the AVs-only zone. This is due to the decreased number of trips using PVs with the expanded AVs-only zone, resulting in fewer vehicles competing selfishly for the shortest paths in the network.

**Validation of model performance regarding data with uncertainty**

In the synthetic demand data created for the case study, two sets of information are generated randomly: departure times and preferences towards CTs or ATs for each group of trips. In reality, trip departure times may fluctuate within a time interval instead of being static. The preference towards CTs or ATs is based on travellers'

perceptions and their personal habits, which may change as well. However, travellers' preferences have a great impact on a city's demand pattern. When the demand pattern changes, it is worthwhile to evaluate the model performance.

Besides the original dataset (denoted as dataset 0), we implemented the proposed solution method using ten different data sets, five of which had departure times that fluctuated by $\pm$ 3 time steps (a total time range of 15 minutes) based on dataset 0 (denoted as datasets 1-5), another five with randomly generated vehicle type preferences (denoted as datasets 6-10). The performance of the solution method with different datasets is displayed in Figure 3.9, in which (a) shows the computational times and (b) shows the maximum number of iterations needed to terminate the algorithm. The computational times are dependent on the required number of iterations and the solution time of the proposed MILP models. When the coverage rate of the AVs-only zone is relatively low (0% and 25%), the algorithm takes fewer iterations but more time to converge compared to other scenarios. This is due to the high demand for PVs at the early stage of the AVs-only zone's expansion. To solve the proposed LLM, a large number of paths are generated resulting in a long solution time of the model in each iteration. On the other hand, the demand for CTs and ATs is relatively small at these stages, leading to a small solution space for PGA. So the algorithm converged easily. When more demand shifts from PVs and CTs to ATs with the expansion of AVs-only zone, the computational time decreases accordingly and more iterations are needed for some datasets because the solution space of PGA is larger even though the solution time for the model is short. When the coverage rate of the AVs-only zone is 100%, no iteration is needed as the fleet sizing problem can be solved by a single-level MILP model.

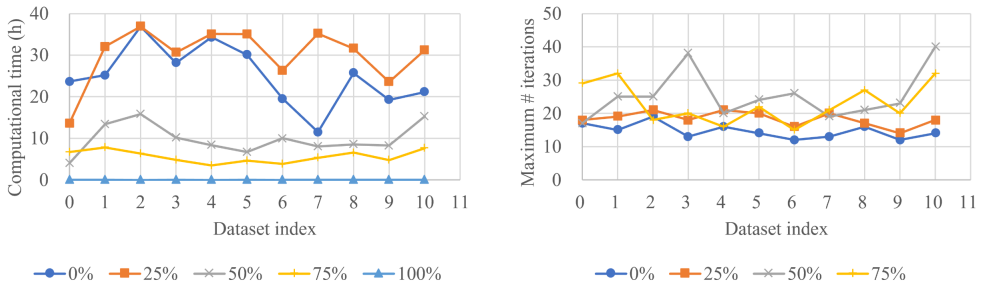



*Figure 3.9: Performance of the solution method with different datasets: (a) Computational time, (b) Maximum number of iterations.*

The computational results are shown in Table 3.12. Analysing the optimisation

results for the first five datasets, we can see a reasonable fluctuation range regarding the best fleets, demonstrating the effectiveness of the proposed solution method. These results provide a TNC with a preliminary insight into choosing the proper fleet sizes considering the randomness of daily trips. A more intuitive suggestion is to take the mean value of all the results. Future research could include a comprehensive stochastic analysis in order to obtain a robust solution. Regarding the results of datasets 6-10, the fleet size fluctuates during the early expansion of the AVs-only zone. This is due to the demand structure change caused by the randomly generated preference toward vehicles. With the increasing coverage rate of the AVs-only zone, more demand will have to be served by ATs (no other option), thereby smoothing the effects of people's preference uncertainty on fleet size decisions.

*Table 3.12: Fleet sizing results for CTs and ATs of different data sets.*

| Coverage rate | 0% | | 25% | | 50% | | 75% | | 100% | |
|---|---|---|---|---|---|---|---|---|---|---|
| Fleet size | Min | Best | Min | Best | Min | Best | Min | Best | Min | Best |
| Dataset 0 | 95, 32 | 95, 43 | 89, 253 | 89, 253 | 27, 550 | 27, 659 | 7, 608 | 7, 714 | 0, 662 | 0, 711 |
| Random departure time | | | | | | | | | | |
| Dataset 1 | 95, 32 | 95, 43 | 86, 257 | 86, 263 | 22, 625 | 22, 639 | 7, 680 | 7, 718 | 0, 666 | 0, 716 |
| Dataset 2 | 92, 28 | 92, 36 | 81, 249 | 81, 251 | 23, 549 | 23, 653 | 7, 578 | 7, 696 | 0, 606 | 0, 693 |
| Dataset 3 | 102, 33 | 102, 41 | 84, 249 | 84, 249 | 23, 625 | 23, 695 | 7, 680 | 7, 782 | 0, 735 | 0, 767 |
| Dataset 4 | 115, 32 | 115, 46 | 85, 274 | 85, 314 | 27, 673 | 27, 673 | 7, 748 | 7, 748 | 0, 784 | 0, 784 |
| Dataset 5 | 100, 32 | 100, 38 | 77, 249 | 77, 249 | 23, 621 | 23, 651 | 7, 676 | 7, 742 | 0, 676 | 0, 723 |
| STD | 8.8, 2.7 | 8.9, 4.0 | 3.7,10.9 | 3.7, 27.9 | 2, 44.4 | 2, 22.0 | 0, 60.7 | 0, 32.5 | 0, 68.3 | 0, 37.7 |
| Random preference towards CTs and ATs | | | | | | | | | | |
| Dataset 6 | 94, 42 | 94, 42 | 89, 250 | 89, 250 | 33, 550 | 33, 653 | 7, 608 | 7, 714 | 0, 662 | 0, 711 |
| Dataset 7 | 79, 57 | 79, 57 | 70, 249 | 70, 249 | 25, 550 | 25, 659 | 11, 608 | 11, 714 | 0, 662 | 0, 711 |
| Dataset 8 | 105, 29 | 105, 40 | 100, 226 | 100, 226 | 30, 555 | 30, 658 | 11, 608 | 11, 714 | 0, 662 | 0, 711 |
| Dataset 9 | 79, 41 | 79, 41 | 77, 260 | 77, 260 | 31, 560 | 31, 663 | 11, 608 | 11, 714 | 0, 662 | 0, 711 |
| Dataset 10 | 80, 42 | 80, 53 | 74, 243 | 74, 243 | 37, 550 | 37, 653 | 11, 608 | 11, 714 | 0, 662 | 0, 711 |
| STD | 11.7, 11.5 | 11.7, 7.8 | 12.3, 12.5 | 12.3, 12.5 | 4.4, 4.5 | 4.4, 4.3 | 1.8, 0 | 1.8, 0 | 0, 0 | 0, 0 |

Looking at the fleet size of CTs in all the tested datasets, we observe again that their minimum feasible fleet size is always the best one. This is because of one significant difference in the cost structure of CTs compared with ATs, which is the drivers' salaries. This observation further corroborates the conclusion drawn in the previous section that the smallest possible fleet size of CTs is always preferable for a TNC in this study.

### Impacts of AVs-only zones

The upgrade of the conventional road networks to AVs-only zones brings inevitable effects on the demand patterns, e-hailing operations, behaviours of travellers, and traffic conditions on road networks. Table 3.11 reveals a clear increase in demand for ATs and a decrease in demand for CTs as HVs (CTs and PVs) are not allowed in most of the network anymore. As a result, the fleet size of ATs increases with the expansion of the

AVs-only zone while that of CTs decreases. When most of the road network is covered by the AVs-only zone, the fleet size of ATs remains stable with the expansion of the AVs-only zone as the usage rate of ATs rises. The total profit of the TNC increases gradually with the expansion of the AVs-only zone.

HVs including both CTs and PVs have to drive outside the AVs-only zone, which results in a longer detour distance and relocation distance in the transition period. Results in Table 3.11 show a significant increase in the relocation distance share of CTs' total travel distance when the coverage rate rises from 0% to 75%. The detour distance share of both CTs and PVs also obviously increases when the coverage rate increases from 0% to 50%. However, with 75% coverage rate of the AVs-only zone, CTs and PVs did not detour. In this case, most of the road links have been converted to AVs-only links. CTs only need to serve a small fraction of the demand in a limited area. Accordingly, the percentage of delivering distance of CTs in total travel distance decreases along with the increase of the percentage of the relocation distance. When ATs are deployed with the best fleet size, there is no significant variation in the percentage of relocation distance and the detour distance. Additionally, the detour only happens to ATs to avoid traffic congestion incurred by competing for the shortest paths. Looking at the results in Table 3.11, there is a slight variation in the percentage of detour distance of ATs which exhibits the same variation tendency as the average delay time per trip of ATs.

In this case study, the AVs-only zone has not necessarily contributed to the reduction of traffic congestion when there is low coverage, even with larger road link capacities. Looking at the total delay time and the average delay time per trip in Table 3.11, these values increase in most cases when the coverage rate goes from 0% to 50%. At the early stage, the benefits of AVs-only zones are not obvious as the demand for ATs is low. However, even at an early stage, the specific delay time of the ATs is lower than those of all HVs because part of the trips are served within the AVs-only zone. In contrast, the congestion effect outside the AVs-only zone increases as the non-automated urban area is further shrunk and vehicles need to compete for the shortest paths. With the expansion of the AVs-only zone, more demand is served by the ATs, and the benefits of the AVs-only zone on decreasing congestion effects begin to unfold. The delay time is largely reduced when most of the urban area is covered by the AVs-only zone. What's more, the total cost for the TNC increases along with the coverage rate of the AVs-only zone up to 50%, as more demand from both CTs and PVs shifts to ATs. Then, it decreases when the coverage rate is 75% and 100% due to the reduced delay penalty and the smaller CT fleet size.

## 3.6 Conclusions and future research

Envisioning the emergence and expansion of AVs-only zones in urban areas, a bi-level framework has been proposed in this chapter to determine the (near) optimal fleet size of CTs and ATs which leads to the maximum profit of a TNC at each stage of a mixed automated and non-automated driving network. At the upper level, the fleet sizing decision of CTs and ATs is made with the aim to maximise the profit of a TNC while satisfying the travel demand. To capture the mixed driving behaviour, an approximated dynamic mixed equilibrium model is proposed at the lower level, in which the respective objective functions of taxis and PVs are combined into one function using a weighted sum approach and the vehicle movements in a morning peak hour of a typical working day are determined. A metaheuristic algorithm PGA is then adopted to solve the bi-level model, which is embedded with a tailored algorithm for solving the LLM.

Computational experiments with the case-study city of Delft show that the (near) optimal solution obtained through the solution method and the minimum fleet sizes of CTs and ATs (minimum feasible fleet to satisfy all the demand) with the expansion of the AVs-only zone can be effectively determined for different datasets with random departure time and random preference towards CTs and ATs. However, the proposed solution approach is hard to apply to a real-size urban network of a metropolis as the computational time can be long and the solution quality cannot be guaranteed. What's more, if a high number of decisions have to be determined in the upper-level model, the solution process can be time-consuming as more iterations are needed until the algorithm converges. Several conclusions can be drawn from the experiments.

Firstly, the minimum fleet size for satisfying the demand is not necessarily the best fleet size for the company's profits. It depends greatly on the cost of the fleet and the drivers. The drivers' salaries, which are one of the highest fleet size-related costs of CTs, have a significant impact on the decision-making process, resulting in that the minimum feasible fleet size of CTs is always their best fleet size for all the tested datasets. Besides, the location and distribution of the parking depots can also influence the fleet size of taxis. TNCs should carefully determine the number of parking depots and locate those depots in areas with high demand to reduce relocation-related costs. Secondly, the existence of AVs-only zones improves transportation efficiency by reducing the congestion effects. But this effect is not obvious at an early stage. To get the best out of using the AVs-only area, governments should consider ways to encourage people to use more AVs at the early stage. Thirdly, the introduction of an AVs-only zone will result in long detours and relocation distances for HVs. Therefore, a proper network design strategy for an AVs-only zone can reduce the negative effects

on HVs, thereby increasing public acceptance of AV-related mobility renovation and the new intelligent infrastructure.

For future research, we recommend studying the following: modelling the fleet sizing problem considering stochastic factors (such as the uncertainty in demand, the fluctuation of traveller's departure time as well as travel times) to make a more robust decision for TNCs; adding travellers' mode choice to describe their preference towards the type of the vehicle; investigating the optimal design strategy of AVs-only zones in a multi-period perspective; and studying the optimal location and distribution of parking depots.

# Appendix

# 3.A Weight determination method

Two weighting coefficients are used in our problem: $\omega$ is used to combine the objectives of taxis and PVs into one single objective function in both LLM and ALLM; and $\lambda$ is used in the objective function of PVs, which is $J^P$, to ensure that the first term of the objective function has absolute priority over the second term. In this section, we introduce methods to determine the value of these two weighting coefficients properly.

## 3.A.1 Determining weighting coefficient $\omega$

The value of the weighting coefficient $\omega$ reflects the priority of the objectives. To give the same priority to the objective of taxis and PVs, the value of $\omega$ need to be properly determined, so that the contribution of both the objective function values can be balanced.

An iterative weight $\omega$ determination method is proposed. Given an initial value to $\omega$, we solve the bi-objective model to obtain the values of $J^T$ and $J^P$ and their contributed values $\omega \cdot J^T$ and $(1-\omega) \cdot J^P$. Given values $J^T$ and $J^P$, a new value of $\omega$ is determined for which $\omega \cdot J^T$ equals $(1-\omega) \cdot J^P$. This procedure is repeated until the relative difference between the contributions of $J^T$ and $J^P$ is less than a predefined small value $\varepsilon$. The pseudo-code is given by Algorithm 3.2 using LLM as an example. The same procedure applies to ALLM by simply substituting $J^P$ for $\hat{J}^P$.

**Algorithm 3.2** Weight $\omega$ determination algorithm

Initialise $\omega$.
Solve the bi-objective optimisation model *LLM*. Calculate the value of $J^T$, $J^P$, $\omega \cdot J^T$ and $(1-\omega) \cdot J^P$.
**while** $\frac{\left|\omega \cdot J^T - (1-\omega) \cdot J^P\right|}{\omega \cdot J^T} \geq \varepsilon$ **do**
  Update the value of $\omega$ by the following equation:
  $\omega := \frac{J^P}{J^T + J^P}$
  Solve the bi-objective optimisation model *LLM* with the newly updated $\omega$.
  Update the value of $J^T$, $J^P$, $\omega \cdot J^T$ and $(1-\omega) \cdot J^P$.
**end while**

## 3.A.2 Determining weighting coefficient $\lambda$

In the LLM, the value of $\lambda$ has to guarantee that the first term in Equation (3.7), which is $\lambda \cdot \sum_{r \in R^{PV}} \frac{K^r}{M^r}$, is always greater than or equal to the second term $\sum_{\pi \in \Pi^r, r \in R^{PV}} \sum_{(i_{t_1}, d_t^r) \in G^{PV}}$

$PF^{r\pi}_{i_{t_1} d^r_t} \left( D^{r\pi} co^{PV} + (t - a^r) ct \right)$. Here, the second term represents the total travel costs of PVs, the worst-case of which is that all PVs take the routes with the maximum cost, which is $\sum_{r \in R^{PV}} K^r \cdot n^r$. To make sure that the first term is greater than or equal to the biggest value of the second term, which is $\lambda \sum_{r \in R^{PV}} \frac{K^r}{M^r} \geq \sum_{r \in R^{PV}} K^r \cdot n^r$, we can let $\lambda = \max(M^r) \cdot \max(n^r)$. In an extreme case, the first term could equal the second term in value. However, to realise that the first term in the objective function has an absolute priority over the second term, meaning that any single unit of increase of the first term is worse than the maximum increase the second term could bring to the objective function value, we let $\lambda = (ub - lb) \cdot \max(M^r) \cdot \max(n^r)$, where $ub$ and $lb$ represent the upper and lower bound values of the second term. The bound values can easily be calculated by knowing the departure time, the latest arrival time and the longest possible path of each group of trips. Here, $(ub - lb)$ represents the maximum increase in the objective function that the second term could bring about. The same method could be used to determine the value of $\lambda$ in ALLM, where $\lambda = (ub - lb) \cdot \max(st^r) \cdot \max(n^r)$.

## 3.B Binary search algorithm

The pseudo-code of the binary search algorithm for finding the lower bound of the fleet sizes is shown in Algorithm 3.3.

## 3.C PGA parameter tuning

Population size is a crucial parameter which influences the algorithm's performance in finding a good solution in the solution space. Using a large population size might increase the possibility of finding an optimal solution but it is not always a good choice when the solution space is small.

We evaluate the performance of PGA with population sizes of 8 and 16, respectively. All other parameters are identical to those listed in Table 3.10. From the results, we see that the best solutions for the four scenarios with 0%, 25%, 50% and 75% AVs-only zones are the same. The only difference is the number of iterations needed, as shown in Table 3.C.1. The algorithm with a population size of 16 is likely to require more iterations than the algorithm with a population size of 8 as more individuals who are less qualified are eligible to be parents, which influences the convergence speed. However, this will not influence the quality of the best solution found in the end.

**Algorithm 3.3** Binary search algorithm for finding the lower bound of fleet sizes

Given the demand for CTs and ATs, calculate the total demand for CTs, ATs as the upper bounds of fleet sizes of CTs, ATs, denoted as $v_{ub0}^{CT}$, $v_{ub0}^{AT}$.
Calculate the maximum number of overlapping travel time intervals for all trips at any point in time as the initial lower bounds of fleet sizes of CTs, ATs, denoted as $v_{lb0}^{CT}$, $v_{lb0}^{AT}$.
Initialise $V^{CT} = V_{lb}^{CT} = v_{lb0}^{CT}$, $V_{ub}^{CT} = v_{ub0}^{CT}$, $V^{AT} = v_{ub0}^{AT}$.
Check the feasibility of the LLM using $v_{lb0}^{CT}$, $v_{ub0}^{AT}$ as input.
**if** LLM is feasible **then**
    The bound of CT's fleet size is $[v_{lb0}^{CT}, v_{ub0}^{CT}]$.
**else**
    **while** $V_{ub}^{CT} - V_{lb}^{CT} \neq 1$ **do**
        $V^{CT} = V_{lb}^{CT} + \left\lfloor (V_{ub}^{CT} - V_{lb}^{CT})/2 \right\rfloor$
        Check the feasibility of the LLM using $V^{CT}$, $V^{AT}$ as input.
        **if** LLM is feasible **then**
            $V_{ub}^{CT} = V^{CT}$
        **else**
            $V_{lb}^{CT} = V^{CT}$
        **end if**
    **end while**
    $V_{lb}^{CT} = V_{ub}^{CT}$
    The bound of CT's fleet size is updated to $[V_{lb}^{CT}, V_{ub0}^{CT}]$.
**end if**
Same process to determine the lower bound of AT fleet size.

*Table 3.C.1: PGA performance for different population sizes.*

| | 8 individuals | 16 individuals |
|---|---|---|
| Number of iterations with 0% coverage rate | 17 | 17 |
| Number of iterations with 25% coverage rate | 18 | 22 |
| Number of iterations with 50% coverage rate | 17 | 28 |
| Number of iterations with 75% coverage rate | 29 | 25 |

## 3.D Similarity threshold selection

A predetermined similarity threshold value $\theta$ is specified in order to reduce the size of the path pool by removing paths that are highly overlapping with other paths. A proper value of this similarity threshold can balance the solution quality and computational efficiency. Our goal is to find a proper similarity threshold value with which the problem can be solved in an acceptable time but also returns a high-quality solution.

Three similarity thresholds of 70%, 80%, and 90% were tested to see how this value affects the objective function value of the found solution and the computational time. Before the optimisation, the fleet sizes of CTs and ATs were given as 95 and 32. This experiment was done in a scenario where the coverage rate of the AVs-only zone is 0%, because this is the scenario with the most demand for PVs. Multiple stopping criteria were set to terminate the algorithm. First, the model could be solved as close to optimality as possible within one hour. After one hour, the algorithm stops either because the MILP gap reached 2%, or because the computational time exceeds ten hours.

The results are shown in Table 3.D.1. Looking at the resulting computational time using these three threshold values, we observe that the algorithm with a 90% threshold value takes the longest time, while its MILP gap value is still greater than those with 70% and 80% threshold values. Regarding the objective function value, the algorithms with a threshold of 80% and 90% have identical values, and the algorithm with a threshold of 70% has a slightly lower value. In conclusion, 80% is an appropriate threshold value in this case as the algorithm could be solved in an acceptable amount of time while guaranteeing the quality of the solution.

*Table 3.D.1: Similarity threshold value.*

| $\theta$ | Objective function value | MIP Gap | Computational time |
|---|---|---|---|
| 70% | 2587.36 | 2.65% | 3600s |
| | 2587.36 | 2.03% | 8075s |
| 80% | 2587.19 | 2.35% | 3600s |
| | 2587.19 | 1.96% | 6953s |
| 90% | 2587.19 | 4.7% | 3600s |
| | 2587.19 | 4.22% | 39600s |

# Chapter 4

# Optimising fleet sizing and management of shared automated vehicle (SAV) services considering endogenous demand, congestion effects, and accept/reject mechanism impacts

In this chapter, we envision a future scenario where non-pooled SAVs replace private cars and provide public on-demand mobility services to satisfy the mobility needs of a city's residents. To help service providers make profitable fleet sizing and management decisions, we develop a mixed-integer non-linear programming model that considers the congestion effects and the mode choice of urban travellers in different income classes, between SAVs and bicycles. In addition, we investigate two types of accept/reject mechanisms (mandatory vs. non-mandatory acceptance) which lead to an endogenously determined acceptance rate that can affect travellers' willingness to use SAV services. The computational challenge posed by the non-linear and non-convex nature of the model is addressed through reformulation and the use of outer-inner approximation methods combined with a breakpoint generation algorithm. We demonstrate the effectiveness of our proposed method in a case study of the city of Delft in The Netherlands, as well as a scaling analysis on three toy networks with various sizes and demand profiles. A sensitivity analysis of key parameters is carried out to assess

system performance.

This chapter is organised as follows: Section 4.1 introduces the background information. The literature on fleet management and demand modelling is reviewed in Section 4.2. Section 4.3 presents the non-linear non-convex mathematical model of the proposed fleet sizing and management problem. In Section 4.4, a detailed description is provided of how to linearise the proposed model, enabling its solvability using state-of-the-art solvers. In Section 4.5, a case study on the city of Delft in the Netherlands is performed. In addition, a scaling analysis is conducted in Section 4.6 to evaluate the model's performance with various network sizes and demand profiles. Section 4.7 gives the main conclusions and provides an outlook on research needs.

## 4.1 Introduction

The idea of replacing private cars with shared mobility services and active modes of transport (walking and cycling) has gained momentum rapidly in recent years. Several main reasons are driving this shift. Firstly, the rising number of private cars has been causing stress in cities, such as the lack of parking for the existing demand, increased traffic congestion, air pollution, energy waste, and traffic accidents between cars and between cars and vulnerable road users. These effects threaten the sustainable development of urban regions. Furthermore, removing private cars within cities can lead to numerous positive impacts on public health, including the reduction of air and noise pollution, heat islands, and the occurrence of injuries (Nieuwenhuijsen & Khreis, 2016). Promoting the use of active modes of transport stands to significantly improve public health by encouraging physical activity. In addition, on-demand mobility systems like Uber and Lyft have gained popularity due to their flexible, seamless, door-to-door services. Consequently, in many cities across the world, the concept of a 'car-free city' is being considered and even adopted. Cities like Hamburg, Oslo, Helsinki, and Madrid have announced their plans to be (partly) private car-free cities, and many cities such as Bogota, Brussels, Chengdu, Copenhagen, and Paris have implemented car-free days (Nieuwenhuijsen & Khreis, 2016).

Given the promising transition towards a transport system without private cars, researchers are exploring future smart mobility solutions for urban implementation, especially considering the high costs associated with traditional public transportation provision. Among these solutions, the utilisation of shared automated vehicles (SAVs) to provide public on-demand mobility services (Spieser et al., 2014; Liang et al., 2020) stands out as one of the most promising. Various benefits have been assessed across different dimensions. As highlighted by Spieser et al. (2014), an automated mobility-on-demand (AMoD) solution has the potential to fulfil the mobility needs of the entire population with approximately one-third of the total number of private cars in operation. Moreover, Fagnant & Kockelman (2014) point out that each SAV could replace around eleven privately owned cars, which brings sizeable energy consumption and greenhouse gas emissions savings. The deployment of SAVs in the transportation system could also lead to reduced parking demand, as revealed by Zhang & Guhathakurta (2017), due to the improved intensity of vehicle utilisation and reduced usage of private vehicles. In the future mobility system, active modes of transport, such as walking and cycling, will continue to be utilised by citizens alongside the provision of public services by SAVs. Walking remains suitable for short-distance trips, while bicycles offer a competitive mode of transport for longer distances due to their numerous advantages.

Bicycles are known for their flexibility, user-friendliness, sustainability, and superior environmental friendliness compared to SAVs, making them an attractive option for individuals seeking to reduce transportation costs. Thus, even with the widespread adoption of SAVs, bicycles are expected to remain prevalent in city centres.

The emergence of such a mobility system will most likely lead to a notable surge in demand for SAV services, consequently boosting the need to enhance the supply capability for on-demand responsive services across time. Simultaneously, the large-scale deployment of SAVs is anticipated to induce substantial shifts in travel behaviour and mode choice, making the future demand profile different from what we have today. Thus, estimating the underlying demand and comprehending the factors that influence the demand profile is important for SAV operators to make the right decisions in fleet sizing and management.

Demand for future SAV services and fleet management will interact with each other. Demand for SAV services has a close relationship with travellers' choice of travel mode behaviour, which is influenced by a variety of factors including price, travel time, service quality, and comfort level associated with a particular mode of transport (Ashkrof et al., 2019; Correia et al., 2019). As a mobility service provider, an SAV operator needs to manage its fleet and provide sufficient service to fulfil the mobility needs of their clients, which could be the entire population if alternative modes are restricted. Generally, decisions within an SAV operation system fall into two main categories: (1) strategic decisions determined prior to service launch (or only questioned between large periods of time), such as fleet sizing, pricing strategy, and service quality level; and (2) operational decisions made and adjusted in real-time in response to incoming requests and the dynamically evolving network status, including trip assignment, vehicle routing, parking, and relocation decisions. This demonstrates the interdependent nature of demand and supply. However, most of the existing studies regarding SAVs assume fixed and known travel demand, which is particularly unsuitable for our problem, given that the demand for the SAV service is currently quite unknown.

Another drawback of assuming travel demand as a known fixed number or as varying linearly with the service level is the oversights of travellers' response to decreased network service levels induced by traffic congestion–a factor that significantly affects demand patterns. This aspect is often disregarded in existing research on fleet sizing and management. Having in mind that the road network is highly congested (higher travel time) and/or that the travel cost is high compared with other transport modes, travellers may adjust their choices. Numerous studies have underscored the profound influence of traffic congestion on travellers' mode preferences. For instance, a preference survey conducted by Chung et al. (2012) in Cheonggyecheon stream in downtown Seoul found a 3.2% decrease in private car usage due to increased congestion,

accompanied by a corresponding 3.6% increase in subway ridership. Additionally, a survey by Tennøy (2010) revealed that 33% of travellers will shift from vehicles to other modes during periods of high congestion. Therefore, congestion plays a critical role in travellers' travel decisions and should be taken into account when predicting demand for SAV services.

In the framework of operations research, most existing research on fleet management problems with traffic congestion and travellers' mode choice utilise simulation-based methods (Gurumurthy et al., 2020; Oh et al., 2020; Pinto et al., 2020; Hörl et al., 2021; Wang et al., 2022b). Although simulation-based methods possess the capability to replicate intricate systems with high levels of detail, they are often time-consuming as a large number of simulations must be executed to evaluate system performance under various scenarios of fleet size and operational rules. Extensive research has been conducted on traffic assignments, aiming to comprehend how traffic congestion influences route choice and travel demand. Nevertheless, traditional traffic assignment only models the flow between origins and destinations, without considering complex planning and operational decision-making for SAV services, such as parking location, relocation strategies, and optimal fleet size. Recent research incorporates traffic assignment into (service) network design problems to evaluate the response of travellers to the (service) network design decisions (Xu et al., 2018b; Pinto et al., 2020; Ye et al., 2021; Cai et al., 2022). Typically framed as bi-level programming models, these problems decide planning decisions at the upper level, and independently model a traffic assignment problem with mode choice at the lower level. However, incorporating the complex operational decisions of SAV services is still challenging.

The centralised control of SAVs provides an opportunity for optimising the planning and complex operational decisions through a single-level model. Unlike human drivers who often prioritise individual route preferences, SAVs can behave cooperatively by following the route guidance from the fleet operator to maximise overall profit. With a shared profit-driven aim, the planning and operational decisions can be addressed at the same level. Therefore, we propose a single-level mathematical programming model from the perspective of an SAV operator to determine the most profitable strategic decisions (fleet size, initial fleet distribution, and service quality level) alongside the operational decisions of a typical day (trip assignment, vehicle routing, parking, and relocations) while considering the congestion effects and the traveller's mode choice between SAVs and active modes of transport depending on their specific income profile. In this study, we take bicycles as the representative of the active modes of transport for the sake of simplicity. We specifically focus on commuting trips during the morning peak hour, and as such, walking is not considered a competitive mode of transportation for SAVs.

We explore two types of accept/reject mechanisms, namely (1) accepting all the requests, and (2) rejecting some requests but the rejection rate will influence travellers' attitudes toward using SAV services, to investigate how service quality levels influence the travel demand and SAV operator supply decisions. This type of accept/reject mechanism is widely considered in dial-a-ride problems when some trips are not profitable or impossible to be picked up or delivered within the desired time windows. However, most papers only consider this mechanism from the service providers' point of view and ignore the effect of service quality level on passengers' willingness to use the service again. A high rejection rate of user requests will lower the probability of a client requesting the service again.

We formulate the proposed problem as a novel mixed integer non-linear non-convex programming model, in which a binary logit model is embedded to describe the travellers' mode choice between SAVs and bikes. Recognising that travellers with different demographic characteristics behave differently in terms of mode choice, we divide the users into three income classes. Each class of travellers perceives the travel utility differently, which is time-dependent, flow-dependent, and path-dependent. The congestion effect is described within the model by dynamically varying travel times with respect to traffic flow. To capture the time-dependent features of traffic flow, we model the vehicle movement through flow variables in a time-space network where the studied time period is discretised and the spatial network is expanded in the time dimension. Different from the traditional vehicle routing problem where each vehicle is tracked individually, we use an aggregated flow which can reduce the number of decision variables in the optimisation model. To facilitate the solution process, we reformulate the model into a mixed-integer linear programming model. Linearisation techniques are proposed to tackle the non-linearity brought by the binary logit model, acceptance rate constraints, and demand calculation constraints. In particular, the outer-inner approximation method together with a breakpoint generation method is proposed to linearise the binary logit model. The breakpoint generation method aims to find the least number of breakpoints with a pre-specified acceptable approximation error. The expression of the maximum approximation error in a specified interval is given.

The main contributions of this chapter are summarised as follows:

- This chapter extends previous work by incorporating travellers' reactions to traffic congestion levels. The fleet management decisions in this context, including the strategic decisions and operational decisions, are constrained by the demand-supply equilibrium modelled through multi-class travellers' mode choice behaviour and are influenced by congestion effects. Thus, the obtained results can

provide more practical and realistic managerial and operation insights for future SAV operators.

- This chapter investigates how accept/reject mechanisms influence the closed demand-supply loop in the context of SAVs. It also highlights the importance of considering the impact of service quality on passengers' willingness to continue using SAV service, emphasising the role of user attitudes and satisfaction in the success and sustainability of SAV systems. To the best of our knowledge, it is still an under-explored topic.
- The proposed flow-based model incorporates three essential elements–mode choice, traffic congestion, and accept/rejection mechanisms–into a single-level optimisation model. This allows us to explore the intricate interactions between these elements, leading to a better understanding of the dynamics in SAV systems.
- The proposed model is validated through comprehensive case studies conducted in the city of Delft, The Netherlands, and three toy networks with various sizes and demand profiles, demonstrating its effectiveness in solving real-world transportation challenges. Additionally, sensitivity analyses on critical parameters are conducted, enabling a thorough assessment of system performance under diverse scenarios.

## 4.2 Literature review

In this chapter, we aim to combine the SAV service fleet sizing and management problem, SAVs congestion modelling, and SAV demand modelling in one optimisation problem. Therefore, from these three aspects, we review the literature to demonstrate the gaps that we have identified as well as search the grounds for the required methodologies for our purpose.

### 4.2.1 Fleet management with congestion effect and travellers' mode choice

A considerable amount of literature has been published on the fleet management problem. The literature includes a wide range of topics such as capacitated vehicle routing problems, vehicle routing problems with time windows, pickup and/or delivery problems, fleet sizing and vehicle routing problems, dial-a-ride transport, etc. This is independent of the vehicles being or not automated. Interested readers can refer to Hyland

& Mahmassani (2017) for the taxonomy of vehicle fleet management problems. More specifically, we classify our problem as an extension of fleet sizing and the vehicle routing problem with travellers' pickup and delivery. The objective of this type of problem is to identify the optimal planning decisions that yield the minimum costs or the maximum profit for the fleet operator, which is constrained by the trip assignment and vehicle operations on the network.

As we mentioned before, most of the existing research on fleet management problems with closed supply-demand loops, the consideration of congestion effect and travellers' mode choice use simulation techniques (Gurumurthy et al., 2020; Oh et al., 2020; Pinto et al., 2020; Hörl et al., 2021; Wang et al., 2022b). Only a handful of studies consider a similar problem using an optimisation-based method or a hybrid approach that combines optimisation with simulation. Wei et al. (2022) study the optimal transit schedules while taking into account the competition with ride-hailing services and traffic congestion. A mixed integer non-linear program is proposed and solved using a bi-level heuristic algorithm including an outer loop and inner loop. The strategic transit scheduling decisions are determined in the outer loop given travellers' mode choice and congestion estimates. The path choice of ride-hailing vehicles and congestion levels are determined in the inner loop in a traffic assignment problem. However, they consider simplified ride-hailing operations by ignoring the relocation of ride-hailing vehicles, and parking decisions. Thus, the congestion effects caused by the re-locations of vehicles can not be captured in the model. Pinto et al. (2020) combine optimisation-based and simulation-based techniques. They propose a bi-level programming model to investigate the integration of the transit network redesign problem and fleet sizing problem for a shared autonomous mobility service. The upper level determines the transit pattern headways and fleet size of SAVs and the lower level describes the combined mode choice–traveller assignment problem. Their approach involves an iterative heuristic procedure where the upper-level problem is solved with a non-linear programming solver and the lower-level problem is solved through agent-based simulation given the decisions made at the upper level.

### 4.2.2 Congestion modelling in fleet management problems

Congestion in road transportation networks has been extensively studied in traffic assignment problems. As one of the major factors influencing the transportation network's performance and decisions related to route choice, congestion effects are increasingly being considered in fleet management problems as well. Liang et al. (2018) envision a future on-demand mobility system where automated vehicles (AVs) serve as taxis to provide mobility services. They take into account the impact of congestion

in determining the optimal trip assignment and dynamic routing of AVs. Expanding on this theme, Liang et al. (2020) delve into a dial-a-ride problem involving ride-sharing in light of the traffic congestion caused by the routing of a large number of AVs. Fan et al. (2022) investigate the heterogeneous fleet sizing and vehicle routing problem for an on-demand mobility service provider envisioning a progressive expansion of AVs-only zones. The congestion effect is incorporated into the model to examine the impacts of the AVs-only zone on travellers' behaviour and network performance.

The congestion effect, measured quantitatively by the variation of travel time as a function of flow, is typically expressed by the well-known Bureau of Public Roads (BPR) function, which is non-linear. Incorporating this non-linear function into a mathematical programming model makes it difficult to solve to optimality. To address this issue, techniques have been proposed including but not limited to: (1) reformulating and linearising the non-linear term (Wang et al., 2015); (2) replacing the BPR function by selecting one from multiple link-traveltime choices at each time point (Van Essen & Correia, 2019); (3) adopting an iterative solution algorithm until the algorithm converges (Correia et al., 2019). In addition to the mathematical programming model, simulation-based methods have also been used as a modelling technique to study the congestion effects on the fleet management problem (Fagnant & Kockelman, 2014; Wang et al., 2022b).

Existing studies on SAV fleet management problems fall short of taking users' preferences and choice behaviours into account, especially behaviours that are influenced by the effect of traffic congestion on travel times. This is an important factor that significantly impacts the mobility pattern and mode preference of travellers, thereby influencing the demand for SAV services and the supply.

### 4.2.3 Demand modelling methods in optimisation

A fundamental assumption used in a large body of literature on fleet management is that demand for all OD pairs is fixed and known in advance (Correia & Van Arem, 2016; Liang et al., 2018; Van Essen & Correia, 2019; Liang et al., 2020; Fan et al., 2022), which does not match with the real world. Assuming travel demand to be constant may lead to unrealistic managerial decisions that result in substantial financial losses for SAV operators. Thus, a more appropriate representation of demand in the fleet management problem is essential.

Demand modelling methods have been widely explored in the existing literature evolving from trip-based models to activity-based models. Trip-based demand modelling representations include but are not limited to the following: (a) elastic demand represented by a simple linear function (Jorge et al., 2015) or a non-linear function

such as an exponential function (Huang & Kockelman, 2020; Huang et al., 2020; Xu & Meng, 2020); (b) probability-based demand representation of discrete choice models, such as the binary logit model (Lu et al., 2021; Guo et al., 2022; Tian et al., 2022), multinomial logit model (Joksimovic et al., 2005; Atasoy et al., 2014; Yang et al., 2022), logit-based chance-constrained model (Dong et al., 2022), mixed logit model (You et al., 2022); (c) disaggregate demand representation of discrete choice models by using simulation-based linearisation (Paneque et al., 2021, 2022); (d) machine-learning methods under a big data context (Wang et al., 2020b). In addition to these trip-based methods, researchers have also developed methods to study activity-based models that focus on the interdependent choice of full daily activity-travel patterns at an individual or household level. These include nested logit models, dynamic discrete choice models (Västberg et al., 2020), machine-learning-based methods (Ren & Chow, 2022), etc.

Among these methods, the discrete choice model is traditionally used to analyse choice behaviours. In this chapter, we incorporate the logit model into our trip-based optimisation problem for the following reasons. Compared with a simple linear function or an exponential function, the logit model is more realistic as it describes the probability of selecting a particular alternative against other alternatives considering a number of factors and their relative importance. It can also take into account the travellers' socioeconomic characteristics such as income level, age, etc. Simulation-based linearisation method is a promising method, but generating a large number of scenarios may bring a big computational burden. Machine learning methods can analyse individual decisions with a higher prediction accuracy but the big data context is missing for the future scenario.

Including the non-linear logit formula in a mathematical model makes the model difficult to solve to optimality due to the non-linear and non-convex formulations. To tackle the non-linearity, researchers proposed many solution methods, such as linearisation algorithm (piece-wise linear function approximation (Wang & Lo, 2010), outer-approximation (Xu et al., 2018a), outer-inner approximation (Liu & Wang, 2015; Guo et al., 2022), heuristic and meta-heuristic (Joksimovic et al., 2005; Lu et al., 2021; Azadeh et al., 2022; Dong et al., 2022; Tian et al., 2022), and simulation (Lou et al., 2011; Paneque et al., 2021; Wang et al., 2022b). Among all of them, one of the most fundamental methods is the piecewise-linear function-based approximation, which aims to find the optimal solution by replacing the non-linear term in the objective function or constraints with a series of piece-wise linear functions. The key idea is to transform the mixed-integer nonlinear programming model into a linear one, and then solve the problem to optimality. A variant of this is the outer/outer-inner approximation method. Instead of replacing the non-linear term with a series of piecewise linear functions, this method specifies the upper bound/the upper and lower bound of

the non-linear term using a series of linear constraints. In this chapter, we adopt the outer-inner approximation method to tackle the computational challenge brought by the logit model.

## 4.3 Problem formulation

In this section, we start by outlining the assumptions of the problem and then provide a detailed description of the mathematical formulation of the proposed model.

### 4.3.1 Assumptions

We envision a future scenario in which cities are only accessible by SAV services and active modes of transportation (e.g. bicycles). Travellers who choose SAVs can request transportation services at any location in a city using SAV service applications on their smartphones by providing trip information. After receiving the trip information, the platform decides whether to accept or reject the trip. Two accept/reject mechanisms are investigated, namely (1) an SAV operator has to accept all the requests, or (2) an SAV operator may reject a trip if it provides no benefits to the company. In this case, those rejected trips will use bicycles. Of course, travellers can choose to use bicycles directly if they perceive that using this mode provides greater travel utility. Once the request is accepted, the platform will match available SAVs with customers and dispatch them to pick up the customers.

Before we can formally introduce the model, we describe the made assumptions.

(a) The proposed model serves a strategic planning purpose. During the study period, the total mobility demand in an urban area is assumed to be constant and known in advance, enabling the SAV operator to make optimal planning decisions. This stands in contrast to real-time SAV operating systems where future demand remains uncertain, necessitating the continuous updating of the optimal operational strategy in response to new incoming demand.

(b) We assume that SAVs and bicycles operate in separate lanes or designated areas, ensuring physical separation and minimal interaction. The flexibility and manoeuvrability of bicycles allow cyclists to easily bypass congested areas and find alternative routes. As a result, our study primarily focuses on analysing the congestion effects caused by the routing of SAVs. It is worth noting the importance of considering interactions between bicycles and AVs, especially in cases where infrastructures do not allow for the separation of traffic (Madigan et al., 2019; Vlakveld et al., 2020; Hulse, 2023). However, we do not extensively delve into this topic within the scope of this

chapter.

(c) We assume that travellers have perfect information about the transportation network status and will make a rational mode choice based on their perceptions of travel time.

(d) We assume that the SAVs in our study are at SAE level 5, and they are capable of driving throughout the entire network without the presence of a human driver.

(e) Neither privately-owned vehicles nor human-driven vehicles are considered as an option.

(f) A traveller will only utilise a single travel mode. Transferring between modes is not considered.

(g) The model considers exogenous fares by taking SAVs. Deciding on the optimal fares is an interesting future direction.

(h) Pooled services are not considered in this study.

To provide a comprehensive understanding of the proposed model, we have illustrated its structure in Figure 4.3.1, which outlines the decisions, input, and output information. Additionally, we have summarised the mathematical notations used in Section 4.3 in Table 4.1 for easy reference.

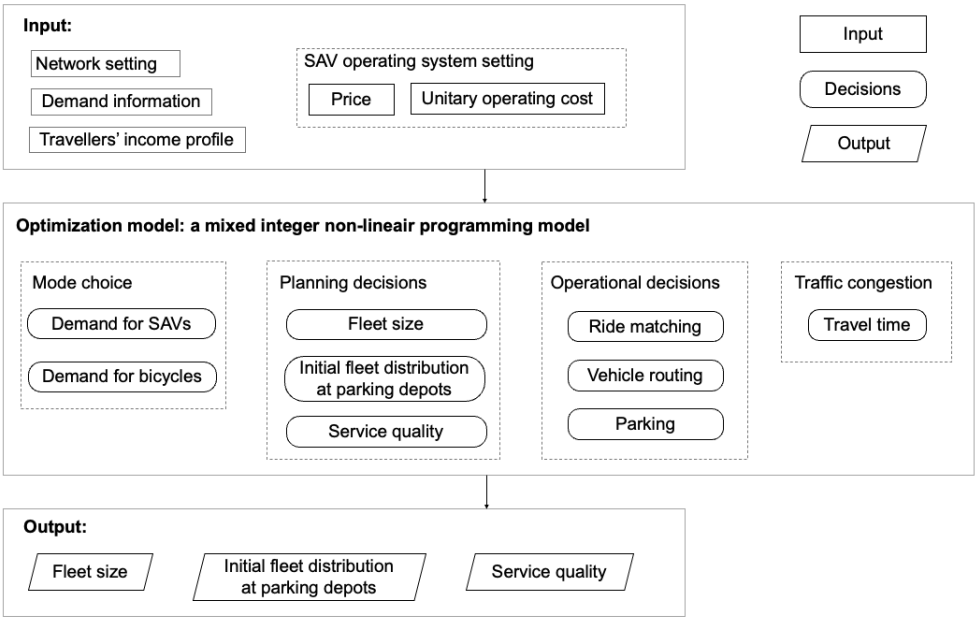



*Figure 4.3.1: Model structure.*

*Table 4.3.1: Notation*

| Notation | Description |
|---|---|
| **Sets** | |
| $T$ | $= \{0, 1, 2, \ldots, \mathscr{T}\}$. Set of time instants in the operation period. |
| $N$ | Set of nodes. |
| $L$ | Set of road links between nodes in set $N$. |
| $G$ | Set of links in the time-space network. |
| $N_P$ | Set of nodes allowing parking for SAVs with $N_P \subseteq N$. |
| $R$ | Set of groups of trips, where each group of trips $r \in R$ has the same origin, destination, departure time, and latest arrival time at the destination. |
| $M$ | Set of travel modes, with the automated vehicles $(AV)$ and bicycles $(B)$ as the two options. |
| **Choice model** | |
| Parameters | |
| $V_B^r$ | Deterministic systematic component of the utility of bicycles for group of trips $r \in R$. |
| $OM_m^r$ | Monetary costs of travellers in group $r \in R$ using mode $m \in M$, in euros. |
| $\beta_0$ | Parameter converting generalised costs into utility, in utility/euro. |
| $\beta_1$ | Parameter converting service rate into utility. |
| $\beta_m^r$ | Travellers' value of travel time in group $r$ using mode $m \in M$, euros/time step. |
| $T_B^r$ | Travel time of using bicycles for trips in group $r \in R$. |
| $n^r$ | Total number of trips for group $r \in R$. |
| Auxiliary variables | |
| $V_{AV}^r$ | Deterministic systematic component of travellers' utility for using an SAV in group $r \in R$. |
| $T_{AV}^r$ | Longest SAVs travel time for group $r \in R$. |
| $P_{AV}^r$ | Probability to choose SAVs for the trips in group $r \in R$. |
| $D_{AV}^r$ | Total number of trips using SAVs in group $r \in R$. |
| **Fleet sizing and management model** | |
| Parameters | |
| $\Delta t$ | Time step. |
| $l_{ij}$ | Length of road link $(i, j) \in L$. |
| $Q_{ij}$ | Capacity of road link $(i, j) \in L$ in vehicles per time step. |
| $t_{ij}^{\max}$ | Maximum travel time by cars on road link $(i, j) \in L$. |
| $t_{ij}^{\min}$ | Minimum travel time by cars on road link $(i, j) \in L$. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of road link $(i, j) \in L$ in vehicles that fit on the road link from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |
| $\alpha$ | Trip service rate when all the requests have to be accepted, %. |
| $o^r$ | Origin node for group of trips $r \in R$. |
| $d^r$ | Destination node for group of trips $r \in R$. |
| $a^r$ | Departure time for group of trips $r \in R$. |
| $b^r$ | Latest arrival time for group of trips $r \in R$. |
| $sd^r$ | Shortest travel distance for group of trips $r \in R$, in kilometres. |
| $st^r$ | Shortest travel time assuming free-flow speed for group of trips $r \in R$, in time steps. |

| | |
|---|---|
| $p^0$ | Initial base fare for using an SAV, in euros/trip. |
| $p$ | Travel distance-related price for using an SAV, in euros/km. |
| $co$ | Unit driving operational cost of an SAV, in euros/km. |
| $cd$ | Penalty for drop-off delay of passengers, in euros/time step. |
| $cf$ | Depreciation cost in one hour for using an SAV, in euros/vehicle. |
| Decision variables | |
| $S^r$ | Total number of trips served by SAVs from group $r$, where $r \in R$. |
| $PF^r_{i_{t_1} j_{t_2}}$ | Passenger flow in the group of trips $r \in R$ served by an SAV in road link $(i,j)$, from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G, a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| Auxiliary variables | |
| $\alpha$ | Trip service rate when some requests can be rejected. |
| $V$ | SAV fleet size. |
| $V_i$ | Initial distribution of SAVs at parking node $i \in N_p$ at the beginning of a day. |
| $E^r_t$ | Total number of passengers in group of trips $r \in R$ arriving at time $t \in T$. |
| $F_{i_{t_1} j_{t_2}}$ | Vehicle flow in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. Note that when $t_1 = 0$, $i \in N_P$, meaning that SAVs have to depart from the parking nodes at the beginning of a day. |
| $W_{i_t}$ | Total number of vehicles parking at node $i \in N_P$ from time instant $t$ to $t+1$, with $t \in T$. |
| $Z^r_t$ | Binary variable with $r \in R, t \in T$ if $a^r + st^r \leq b^r$. |
| $X_{i_{t_1} j_{t_2}}$ | Binary variable which is 1 when any vehicle travels in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. |
| $A^r_t$ | Binary variable which is 1 when at least one trip in group $r \in R$ arrives at time $t \in T$, and 0 otherwise. |

### 4.3.2 Network representation

To capture the endogenous dynamic traffic congestion caused by the large-scale deployment of SAVs, we utilise a time-space network. A time-space network is a time-expanded version of the directed physical network $(N,L)$, where $N$ and $L$ denote the set of nodes and road links, respectively. The time dimension is discretised into $\mathscr{T}$ periods, with each period having a duration $\Delta t$, referred to as the time step, hereinafter. Consequently, an index set of time periods $T = \{0,1,2,\ldots,\mathscr{T}\}$ is defined within the study horizon. At each time instant $t \in T$, the network is replicated, thus multiple networks are defined along with the time period, as shown in Figure 4.3.2. We define set $G$ to denote the set of links in the time-space network.

Different from the traditional physical network, the status of vehicles on the time-space network is described by both the action of the vehicles and the time those actions take. Thus, vehicles either move with passengers or relocate without passengers on links $(i_{t_1}, j_{t_2}) \in G$, representing the flow departing from node $i \in N$ to node $j \in N$ from time instant $t_1 \in T$ to time instant $t_2 \in T$, or park at node $i \in N_p$ from time instant $t$ to $t+1$ where $t \in T$. $N_p$ denotes the subset of nodes $N$ allowing parking

for SAVs. Here, the parking depots are restricted parking areas spread across the city that are provided by the SAV operator for their vehicles; therefore, parking is only permitted at certain designated nodes. Figure 4.3.2 provides an illustrative example of vehicle movements within a time-space network, depicting passenger deliveries using solid lines, relocation movements without passengers heading to parking stations or passengers' pick-up/drop-off locations with dashed lines, and parking states displayed by dotted lines. In this small physical network, we assume a uniform travel time of 1 time step for each link.

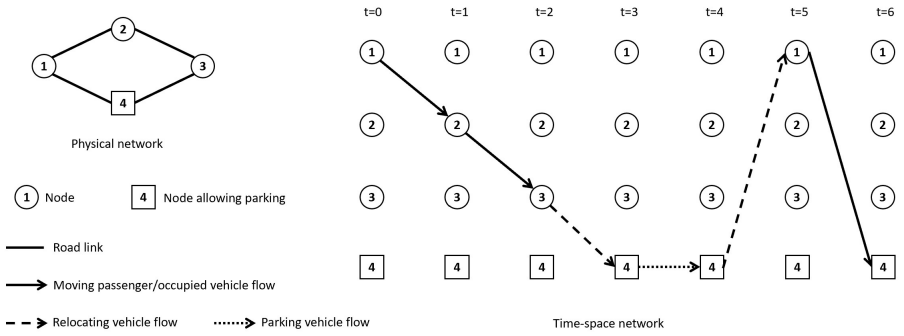



*Figure 4.3.2: Illustration of the physical network, the time-space network, and vehicle movements over the time-space network.*

Several physical attributes related to the road links are defined, such as the length of road link $(i, j) \in L$ denoted as $l_{ij}$, the capacity of road link $(i, j) \in L$ per time step denoted as $Q_{ij}$, and the maximum and minimum travel time of link $(i, j) \in L$ denoted as $t_{ij}^{\max}$ and $t_{ij}^{\min}$, respectively. Given the maximum and minimum travel time of link $(i, j) \in L$, we can further shrink the size of set $G$ by only including the possible time choices instead of doing a complete enumeration for all the time instants. It means that we only include link $(i_{t_1}, j_{t_2})$ where $t_1 + t_{ij}^{\min} \leq t_2 \leq t_1 + t_{ij}^{\max}$.

Please note that when using a time-space network framework, travel time is represented in an integer number of time periods. However, this representation does not imply that the actual travel time must be an integer value. The integer value corresponds to the index of the time period within the study horizon. The actual value of travel time depends on the chosen time step, which may or may not be an integer. However, we do recognise that the time step sets the precision of travel times where the maximum precision is always limited by the duration of the time step.

### 4.3.3 Demand representation and choice modelling

We consider the peak hour of a typical workday in an urban area, where congestion occurs which influences travellers' mode choices. Instead of tracking each trip individually, demand that shares the same travel information is aggregated into a group. We introduce set $R$ as the set of groups of trips, where each group of trips $r \in R$ has the same origin $o^r \in N$, destination $d^r \in N$, departure time $a^r \in T$, latest arrival time $b^r \in T$, shortest travel distance $sd^r$ in kilometres, shortest travel time $st^r$, and the same income level in euro/time step. Note that trips within a group only have the same departure time, latest arrival time, and shortest travel time based on time periods, not based on the real times in seconds. Trips from one group are allowed to use different transport modes and the total number of trips in group $r \in R$ is denoted as $n^r$. We denote $M$ as the set of travel modes, with the shared automated vehicles ($AV$) and bicycles ($B$) as the two options.

The mode choices for travellers in each group are analysed using discrete choice modelling within the framework of random utility maximisation theory. To measure the willingness or preference for using a certain type of travel mode, the utility of each mode is calculated. Besides that, three income levels for travellers are considered: high income, middle income, and low income. The income level determines the value of travel time (VOTT) of travellers, which has a direct impact on their perceived utility for using a particular travel mode and consequently influences their mode choice. The VOTT of using travel mode $m \in M$ for travellers in the same group $r \in R$ is assumed to be the same, which is denoted by $\beta_m^r$ in euro/time step. The three classes are not explicitly defined in the model but are implicitly included in the travel information of each group of trips.

However, the utility is not known with certainty due to factors such as unobserved variation among travellers, unobserved attributes of the alternatives, and perception errors of travellers (Ben-Akiva et al., 1985). Therefore, the utility of mode $m \in M$ for trips in the group $r \in R$, denoted as $U_m^r$, is treated as a random variable. It consists of a deterministic systematic component $V_m^r$, which is the observable utility of mode $m \in M$ for trips in the group $r \in R$, and a random component $\varepsilon_m^r$, which is the unobservable component of the utility.

$$U_m^r = V_m^r + \varepsilon_m^r, \qquad \forall r \in R, m \in M \tag{4.1}$$

The deterministic term $V_{AV}^r$ of the utility for using an SAV for group $r \in R$ depends on the generalised cost of using SAVs and travellers' satisfaction towards SAV services. To be more specific, the generalised cost of using SAVs for group of trips $r \in R$ is calculated as the linear sum of the fare of travellers $OM_{AV}^r$ and the travel time-related

cost of the journey $\beta^r_{AV} T^r_{AV}$. The fare $OM^r_{AV}$ of using the SAV service for travellers in group $r \in R$ depends on the pricing strategy of SAV operators and the travel distance of a trip. The perceived travel time $T^r_{AV}$ is affected by the dynamically varying traffic congestion, which can be determined endogenously by solving the optimisation model. $\beta_0$ is a parameter that converts generalised costs into utility, expressed as utility/euro, which indicates the sensitivity of travellers to the change in the monetary costs. Travellers' satisfaction with the SAV service depends on the trip service rate $\alpha$. When the SAV operator is not allowed to reject any trips, the trip service rate equals 1 (100%). However, if some trips are rejected by the SAV operator, the trip service rate will be less than 1, which will decrease the traveller's satisfaction with the SAV service. $\beta_1$ is the parameter that describes travellers' satisfaction with the service rate.

$$V^r_{AV} = -\beta_0(OM^r_{AV} + \beta^r_{AV} T^r_{AV}) - \beta_1(1-\alpha), \qquad \forall r \in R \tag{4.2}$$

Alternatively, the deterministic term $V^r_B$ of the utility for using a bicycle for group $r \in R$ is calculated based on the monetary cost of using bicycles, as shown in Equations (4.3). The monetary cost $OM^r_B$ for group $r \in R$ is the bicycle's depreciation cost which is calculated by dividing the bicycle's purchase price by its service life. We assume that the travel time $T^r_B$ for group $r \in R$ is a constant, as the congestion in motor lanes will not affect the travel time of bicycles.

$$V^r_B = -\beta_0(OM^r_B + \beta^r_B T^r_B), \qquad \forall r \in R \tag{4.3}$$

In Equations (4.1), $\varepsilon^r_m$ is the error between the actual utility and the systematic utility of mode $m \in M$ for trips in group $r \in R$, which can be viewed as the part of utility that is unknown to the analyst. Assuming these error terms are all independently and identically Gumbel distributed, we can compute the probability of choosing SAVs against bicycles in group $r \in R$, denoted as continuous variables $P^r_{AV}$, by using a binary logit model shown in Equations (4.4).

$$P^r_{AV} = \frac{e^{V^r_{AV}}}{e^{V^r_{AV}} + e^{V^r_B}}, \qquad \forall r \in R \tag{4.4}$$

We introduce integer variables $D^r_{AV}$ to represent the demand for SAVs for group $r \in R$, which can be calculated using the total number of trips $n^r$ in group $r \in R$ multiplied by their probability of choosing SAVs. Then, we round this value to the nearest integer using a floor function as shown in Equations (4.5).

$$D^r_{AV} = \lfloor n^r P^r_{AV} + 0.5 \rfloor, \qquad \forall r \in R \tag{4.5}$$

### 4.3.4 Fleet sizing and management for SAV operators

In this section, we develop the base formulation for an SAV operator to manage the SAV fleet. Three tiers of decisions are made in this model: (1) at the strategic level: the overall SAV fleet size, the initial fleet distribution at the beginning of a day and the service quality level; (2) at the operational level: the assignment of passengers to SAVs, vehicle routes determination, parking and relocation decisions; and (3) the travel time on each road link.

For each group $r \in R$, the total number of trips served by SAVs is specified by integer variables $S^r$. Therefore, the relationship between the total number of served trips, the total demand and the service rate can be described by Constraint (4.6). It should be noted that when an SAV operator must accept all the requests to maintain a high level of service quality, the parameter $\alpha$ is set to 1. When an SAV operator can reject those requests that bring no profits for the company, $\alpha$ is defined as a continuous variable where $\alpha \in [0,1]$ and its value is determined endogenously by solving the model. As a result, Constraint (4.6) becomes a non-linear constraint.

$$\alpha \sum_{r \in R} D^r_{AV} = \sum_{r \in R} S^r \tag{4.6}$$

In addition, for each of the group $r \in R$, the number of served trips $S^r$ should be less than the demand for SAVs $D^r_{AV}$.

$$S^r \leq D^r_{AV}, \qquad \forall r \in R \tag{4.7}$$

The movement of vehicles is modelled as flow circulating on the time-space network. We introduce integer variables $PF^r_{i_{t_1} j_{t_2}}$ to represent the passenger flow in the group of trips $r \in R$ served by an SAV in road link $(i,j)$, from time instant $t_1$ to $t_2$. These variables are only defined for $(i_{t_1}, j_{t_2}) \in G, a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. Passengers in the same group $r \in R$ are picked up by the SAVs at the origin node $o^r$ at the departure time $a^r$ where $r \in R$. This is ensured by Constraints (4.8). In this study, we do not model the explicit waiting time from the passenger's perspective. We focus on a strategic planning problem, assuming that all the trip information is available in advance. This enables the service operator to make optimal planning decisions. Our primary focus is to determine the required number of vehicles to ensure that travellers depart at their desired times. In addition, this analysis concentrates on two urban travel modes: SAVs and bicycles. Our expectation is for the SAV operator to deliver a high-quality service, with a short waiting time for travellers before pick-up. Nevertheless, we do acknowledge that our approach possesses limitations in scenarios where passengers make their requests without sufficient advance notice and

demand immediate vehicle availability. In such cases, ignoring passenger waiting time could impact strategic decision-making. However, the explicit inclusion of waiting time through modelling could increase the model's complexity, presenting challenges in achieving optimal solutions. As a result, to strike the balance between computational complexity and accuracy, we choose not to model the waiting time explicitly.

$$S^r = \sum_{j_t | \left(o^r_{a^r}, j_t\right) \in G} PF^r_{o^r_{a^r} j_t}, \qquad \forall r \in R \tag{4.8}$$

When delivering passengers to their destination $d^r$, SAVs serving the same group of trips $r \in R$ are allowed to take alternative routes to evenly distribute the flow and alleviate the network burden. For this reason, SAVs may arrive at the destination at different times, but not later than the user-specified latest arrival time $b^r$. To describe this, integer variables $E^r_t$ are defined to represent the total number of passengers in group of trips $r \in R$ arriving at time $t \in T$ with $a^r + st^r \leq t \leq b^r$. Constraints (4.9) and (4.10) ensure that the number of served trips in group $r \in R$ is equal to the number of trips arriving at destination $d^r$ at different times.

$$S^r = \sum_{t \in T | a^r + st^r \leq t \leq b^r} E^r_t, \qquad \forall r \in R \tag{4.9}$$

$$E^r_t = \sum_{i_{t_1} | \left(i_{t_1}, d^r_t\right) \in G} PF^r_{i_{t_1} d^r_t}, \qquad \forall r \in R, a^r + st^r \leq t \leq b^r \tag{4.10}$$

Constraints (4.11) and (4.12) guarantee that the destination node $d^r$ and the origin node $o^r$, respectively, of the group of trips $r \in R$ will only be visited once during client delivery. It indicates that passengers will be dropped off at the destination node the first time the SAV arrives there, and SAVs will not return to the origin node after departure.

$$\sum_{j_{t_1} | \left(d^r_t, j_{t_1}\right) \in G} PF^r_{d^r_t j_{t_1}} = 0, \qquad \forall r \in R, a^r \leq t \leq b^r \tag{4.11}$$

$$\sum_{i_{t_1} | \left(i_{t_1}, o^r_t\right) \in G} PF^r_{i_{t_1} o^r_t} = 0, \qquad \forall r \in R, a^r \leq t \leq b^r \tag{4.12}$$

The next constraints describe the passenger flow conservation at any nodes $i \in N$ in the network except the origin node $o^r$ and destination node $d^r$ for the group of trips $r \in R$.

$$\sum_{j_{t_1} | \left(j_{t_1}, i_t\right) \in G} PF^r_{j_{t_1} i_t} = \sum_{j_{t_2} | \left(i_t, j_{t_2}\right) \in G} PF^r_{i_t j_{t_2}}, \qquad \forall r \in R, a^r < t < b^r, i \in N, i \neq o^r, i \neq d^r \tag{4.13}$$

On the road, SAVs have three statuses: (1) transporting a passenger; (2) driving empty to pick up the next passenger or driving to a parking depot; and (3) being parked at a depot. We introduce continuous variables $F_{i_{t_1} j_{t_2}}$ to describe the vehicle flow (the number of SAVs) in road link $(i, j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. Note that when $t_1 = 0$, only links with $i \in N_P$ are defined, meaning that SAVs have to depart from the parking nodes at the beginning of the day. In addition, continuous variables $W_{i_t}$ are defined to represent the total number of SAVs parking at node $i \in N_P$ from time instant $t$ to $t+1$, with $t \in T$. Note that there is no need to define $F_{i_{t_1} j_{t_2}}$ and $W_{i_t}$ as integer variables explicitly. The integrality requirement for these variables is implicitly satisfied through Constraints (4.14)-(4.17), which will be explained later.

No matter in which status, the vehicle flows $F_{i_{t_1} j_{t_2}}$ in the time-space network should always be greater than the passenger flow $\sum_{r \in R} PF^r_{i_{t_1} j_{t_2}}$ to satisfy the mobility need, as indicated by Constraints (4.14).

$$\sum_{r \in R} PF^r_{i_{t_1} j_{t_2}} \leq F_{i_{t_1} j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{4.14}$$

The next constraints describe the flow conservation rules applied to SAVs' circulation at both normal and parking nodes.

$$\sum_{j_{t_1} | (j_{t_1}, i_t) \in G, t_1 < t} F_{j_{t_1} i_t} = \sum_{j_{t_2} | (i_t, j_{t_2}) \in G, t < t_2} F_{i_t j_{t_2}}, \qquad \forall i \in N \setminus N_P, 0 < t < \mathscr{T} \tag{4.15}$$

$$\sum_{j_{t_1} | (j_{t_1}, i_t) \in G, t_1 < t} F_{j_{t_1} i_t} + W_{i_{t-1}} = \sum_{j_{t_2} | (i_t, j_{t_2}) \in G, t < t_2} F_{i_t j_{t_2}} + W_{i_t}, \qquad \forall i \in N_P, 0 < t < \mathscr{T} \tag{4.16}$$

It is worth mentioning that the vehicle flow $F_{i_{t_1} j_{t_2}}$ is associated with a vehicle flow-related cost in the objective function, which is minimised. Further details about the objective function can be found in Section 4.3.6. Consequently, the minimum number of vehicles required to transport passengers in link $(i_{t_1}, j_{t_2})$ equals the total number of passengers $\sum_{r \in R} PF^r_{i_{t_1} j_{t_2}}$ according to Constraints (4.14), which is an integer. Besides the occupied vehicle flow, $F_{i_{t_1} j_{t_2}}$ also includes the empty relocating vehicle flow. These empty vehicles are either driving after delivering the passengers or are on their way to pick up new passengers. According to the vehicle conservation constraints, these flows may be integer values as well.

The initial distribution of SAVs at parking node $i \in N_p$ at the beginning of a day is denoted by integer variables $V_i$. At the beginning of the optimisation period, SAVs either depart from the parking depots to pick up passengers or park at the parking node waiting for the task given by the SAV operator, as described in Constraints (4.17).

Given that both $F_{i_0,j_t}$ and $V_i$ take integer values, $W_{i_0}$ will also take integer values.

$$\sum_{j_t|(i_0,j_t)\in G} F_{i_0 j_t} + W_{i_0} = V_i, \qquad \forall i \in N_p \tag{4.17}$$

In addition, the sum of the initial fleet distributed at the parking nodes gives the overall SAV fleet size, denoted as integer variable $V$, as shown in Constraint (4.18).

$$\sum_{i\in N_p} V_i = V \tag{4.18}$$

To specify the longest travel time of trips in group $r \in R$, we introduce binary variables $A_t^r$ which are 1 when at least one trip in group $r \in R$ arrives at time $t \in T$, and 0 otherwise. Constraints (4.19) specify the arrival times of trips in group $r \in R$. Then, Constraints (4.20)-(4.22) calculate the longest travel time experienced by trips in group $r \in R$. Constraints (4.20) impose a lower bound to the longest travel time of trips in group $r \in R$ meaning that it has to be bigger than or equal to all of the different travel times experienced by travellers. Constraints (4.21) and (4.22) impose an upper bound to the longest travel time meaning that it has to be less than or equal to the longest travel time among all of the different travel times experienced by travellers in group $r \in R$. We define binary variables $Z_t^r$ and impose that $\sum_{t|a^r+st^r\le t\le b^r} Z_t^r$ equals 1 to ensure that variable $T_{AV}^r$ can only take one value which is the longest travel time of travellers using SAVs in group $r \in R$. $\mathscr{M}$ in Constraints (4.21) is a sufficiently large number.

$$\frac{E_t^r}{n^r} \le A_t^r \le E_t^r, \qquad \forall r \in R, a^r + st^r \le t \le b^r \tag{4.19}$$

$$T_{AV}^r \ge A_t^r (t - a^r), \qquad \forall r \in R, a^r + st^r \le t \le b^r \tag{4.20}$$

$$T_{AV}^r \le A_t^r (t - a^r) + \mathscr{M}(1 - Z_t^r), \qquad \forall r \in R, a^r + st^r \le t \le b^r \tag{4.21}$$

$$\sum_{t|a^r+st^r\le t\le b^r} Z_t^r = 1, \qquad \forall r \in R \tag{4.22}$$

### 4.3.5 Traffic congestion

We include traffic congestion in the model by introducing flow-dependent travel time. It is important to note that only the flow of SAVs contributes to traffic congestion, as this study does not explore the mixed flow interaction between SAVs and bicycles. According to the BPR function (Dafermos & Sparrow, 1969), the travel time of traversing a road link has a non-linear relationship with the vehicle flow on this link: $t = t_0 \left(1 + a\left(\frac{F}{Q}\right)^b\right)$. Here, $a$ and $b$ are parameters; $t_0$ represents the minimum travel

time, $F$ denotes the vehicle flow, and $Q$ represents the link capacity. However, involving the non-linear BPR function makes the solving process more difficult. Therefore, we adopt the method proposed by Van Essen & Correia (2019) to replace the non-linear travel time calculation with multiple link-time-capacity choices. They use the concept of spatial link capacity $C_{i_{t_1} j_{t_2}}$ of a certain link $(i, j) \in L$ within a travel time slot between $t_1 \in T$ to $t_2 \in T$. The spatial link capacities can be calculated before the optimisation using the following equation.

$$C_{i_{t_1} j_{t_2}} = (t_2 - t_1) Q_{ij} \left( \frac{1}{a} \left( \frac{t_2 - t_1}{t_{ij}^{\min}} - 1 \right) \right)^{\frac{1}{b}} \tag{4.23}$$

We add 0.5 to $t_2$ when $t_2 - t_1$ equals $t_{ij}^{\min}$ to prevent the value of $C_{i_{t_1} j_{t_2}}$ from being zero. Among all the link-time-capacity choices, only one can be selected, meaning that there is a unique travel time for traversing road link $(i, j)$ at a time instant $t_1 \in T$. This is described in Constraints (4.24) by making use of binary variables $X_{i_{t_1} j_{t_2}}$ which are 1 when any vehicle travels in road link $(i, j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. Note that using a large set of binary variables $X_{i_{t_1} j_{t_2}}$ in the time-space network may increase the complexity of solving the model (Kaufman et al., 1998). Specifying the set of binary variables requires the enumeration for each road link $(i, j) \in L$ at each time instant $t_1 \in T$ and $t_2 \in T$. However, in our case, given the maximum and minimum travel time of link $(i, j) \in L$, we only define binary variables for each link $(i, j)$ from time instant $t_1 \in T$ to $t_2 \in T$ if $t_1 + t_{ij}^{\min} \leq t_2 \leq t_1 + t_{ij}^{\max}$. This reduces the number of binary variables required.

$$\sum_{t_1 | (i_t, j_{t_1}) \in G} X_{i_t j_{t_1}} \leq 1, \qquad \forall (i, j) \in L, t \in T \tag{4.24}$$

Constraints (4.25) require that the total flow on road link $(i, j)$ from time instant $t_1$ to time instant $t_2$ never exceeds its corresponding spatial link capacity. Given that only one specific travel time will be chosen defined by Constraints (4.24), many flow variables $F_{i_{t_1} j_{t_2}}$ are imposed to zero.

$$F_{i_{t_1} j_{t_2}} \leq \left\lfloor C_{i_{t_1} j_{t_2}} \right\rfloor X_{i_{t_1} j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{4.25}$$

Vehicles that enter the road link first will leave the road link. This is known as the first-in-first-out (FIFO) rule, described by Constraints (4.26). These constraints only apply to time instant $t_1$ and $t_2$ when $t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min}$. Otherwise, if $t_2 + t_{ij}^{\min} > t_1 + t_{ij}^{\max}$, there is no need to impose FIFO rule as vehicles entering the road

link $(i, j)$ at time instant $t_1$ with the longest travel time have left the link before any vehicles entering the road link $(i, j)$ at a later time instant $t_2$ despite travelling with the shortest travel time.

$$\begin{gathered} t_1 + \sum_{t \in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \sum_{t \in T} X_{i_{t_2} j_t} (t - t_2) + \mathscr{M} \left( 1 - \sum_{t \in T} X_{i_{t_2} j_t} \right), \\ \forall (i, j) \in L, t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min} \end{gathered} \tag{4.26}$$

### 4.3.6 Objective function

With the purpose of maximising the total profit of the SAV operator, we define the objective function as Equation (4.27), which comprises the total revenue paid by the service users and the total costs of operating the whole system. Service users have to pay two types of fares for using an SAV, an initial fixed base fare $p^0$ in euros, and a distance-related price $p$ in euros per kilometre. The distance-related price is charged based on the shortest travel distance $sd^r$ of the trip $r \in R$ instead of the actual travel distance to avoid the unnecessary detours of SAVs in order to earn extra profits.

The total costs include the following: (1) the total depreciation costs of the SAV fleet in the system, which is calculated as the unit depreciation cost in euros per vehicle, denoted as $cf$, multiplied by the total fleet size. Here, $cf$ is calculated as the vehicle's purchase price divided by its service life span; (2) the total operational costs including fuel, maintenance and insurance cost, which are calculated by multiplying the total distance of all the SAVs by the unit operational cost $co$ in euros per kilometre; note that the total travel distance includes both the deliver distance with clients and the relocation distance without clients; (3) the penalty for the late drop-off of the client, calculated by multiplying the delay cost $cd$ in euros per time step by the difference between the actual riding time of clients and the shortest travel time.

$$\max \sum_{r \in R} OM_{AV}^r S^r - cf \cdot V - co \left( \sum_{(i_{t_1}, j_{t_2}) \in G} l_{ij} F_{i_{t_1} j_{t_2}} \right) - cd \sum_{r \in R} \left( \sum_{t \in T} t E_t^r - a^r S^r - st^r S^r \right) \tag{4.27}$$

where

$$OM_{AV}^r = p^0 + sd^r p, \qquad \forall r \in R. \tag{4.28}$$

## 4.4 Problem linearisation

The model proposed in Section 4.3 is a non-linear mixed integer programming model because of the exponential terms in binary logit model in Equations (4.4), the floor function to calculate the SAV demand in Equations (4.5) and the quadratic term to determine the acceptance rate in Constraint (4.6). To facilitate the solution process, we propose methods to linearise these non-linear equations and constraints, thereby transforming the model into a mixed integer linear programming model. In addition, we determine the most appropriate value for $\mathscr{M}$ used in Constraints (4.21) and (4.26) to get a tighter formulation of the proposed model.

### 4.4.1 Linearisation of the binary logit model

In Section 4.4.1, we first reformulate the binary logit model in Equations (4.4) with logarithmic functions which are still non-linear. Then, we adopt the outer-inner approximation method to linearise the logarithmic functions. Details of this method are described in Section 4.4.1. To use this method, a set of breakpoints needs to be prespecified before the optimisation. Thus, in Section 4.4.1, we propose a breakpoint determination method to find the fewest breakpoints while guaranteeing a certain level of approximation accuracy.

#### Model reformulation

We firstly rewrite Equations (4.4) as:

$$\frac{P^r_{AV}}{1-P^r_{AV}} = \frac{e^{V^r_{AV}}}{e^{V^r_B}}, \qquad \forall r \in R. \tag{4.29}$$

Then, we take the logarithm of both sides of Equations (4.29) to have the following equation:

$$\ln P^r_{AV} - \ln(1-P^r_{AV}) = V^r_{AV} - V^r_B, \qquad \forall r \in R. \tag{4.30}$$

By defining new variables $LN^r_{AV}$ and $LN^r_B$, we can further simplify the equation by having the following:

$$LN^r_{AV} = \ln P^r_{AV}, \qquad \forall r \in R, \tag{4.31}$$

$$LN^r_B = \ln(1-P^r_{AV}), \qquad \forall r \in R, \tag{4.32}$$

$$LN_{AV}^{r} - LN_{B}^{r} = V_{AV}^{r} - V_{B}^{r}, \qquad \forall r \in R. \tag{4.33}$$

**Linearisation of logarithmic functions: outer-inner approximation-based linear programming relaxation**

We adopt the outer-inner approximation method (Wang et al., 2015; Guo et al., 2022) to linearise the logarithmic term in Constraints (4.31) and (4.32). The logarithmic function can be relaxed to a set of linear constraints that give the upper and the lower bound of the original logarithmic function (see Figure 4.4.1). The procedure for linearising Constraints (4.31) and (4.32) are the same. For the sake of simplicity, we only take Constraints (4.31) as an example.

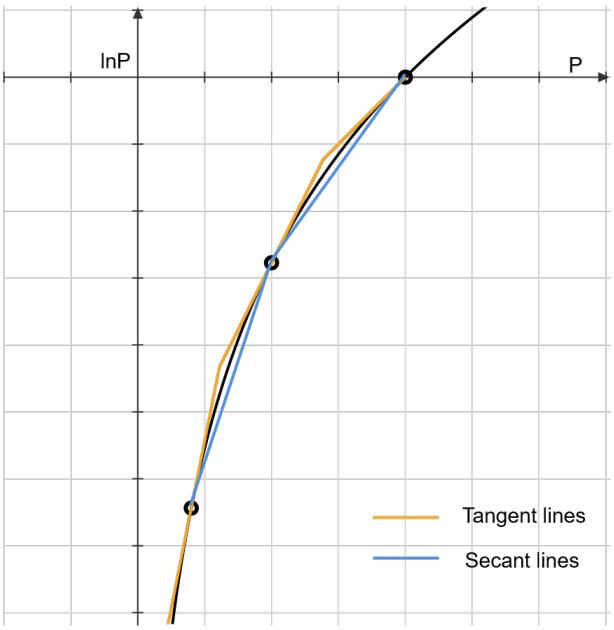



*Figure 4.4.1: Outer-inner approximation.*

The original logarithmic function is divided into $\mathscr{K}-1$ segments by $\mathscr{K}$ predetermined breakpoints ($\mathscr{K}$ is the number of breakpoints). For each segment, the tangent lines at the two breakpoints and the secant line between the breakpoints serve as the upper bound and the lower bound of the real logarithm function, respectively. Note that the breakpoints can be distributed non-uniformly to minimise the approximation error. We introduce an index set for breakpoints, denoted by $K = \{1, 2, \ldots, k, \ldots, \mathscr{K}\}$.

Each breakpoint $k \in K$ has coordinates $(u^k, \ln u^k)$ where $u^k \in [0,1]$. The segment between two adjacent breakpoints $k \in K$ and $k+1 \in K$ is denoted by $[u^k, u^{k+1}]$. If the value of variable $P^r_{AV}$ falls within the interval $[u^k, u^{k+1}]$, we say that $[u^k, u^{k+1}]$ is an active interval. A binary variable $\lambda^k_r$ is defined for $k \in \{1,2,\ldots,k,\ldots,\mathscr{K}-1\}$, $r \in R$ to indicate whether or not an interval $[u^k, u^{k+1}]$ is active for group $r \in R$.

A set of constraints is introduced to describe this outer-inner approximation. Constraints (4.34) describe the tangent lines at each breakpoint, which serve as the outer approximation of the logarithmic function.

$$LN^r_{AV} \leq \frac{1}{u^k} P^r_{AV} + \ln u^k - 1, \qquad \forall r \in R, k \in K \tag{4.34}$$

Constraints (4.35)-(4.41) describe the inner approximation of the logarithmic function. The value of variables $LN^r_{AV}$ and $P^r_{AV}$ can be represented by the convex combination of the coordinates of two consecutive breakpoints, where continuous variables $\theta^k_r$ are defined to represent the convex combination coefficient for breakpoint $k \in K$ for group of trips $r \in R$, as shown in Constraints (4.35) and (4.36).

$$LN^r_{AV} \geq \sum_{k=1}^{\mathscr{K}} \theta^k_r \ln u^k, \qquad \forall r \in R \tag{4.35}$$

$$P^r_{AV} = \sum_{k=1}^{\mathscr{K}} \theta^k_r u^k, \qquad \forall r \in R \tag{4.36}$$

The summation of coefficient $\theta^k_r$ has to be one, according to the convexity Constraints (4.37).

$$\sum_{k=1}^{\mathscr{K}} \theta^k_r = 1, \qquad \forall r \in R \tag{4.37}$$

There exists only one active interval, meaning that the value of $P^r_{AV}$ and $LN^r_{AV}$ can only fall into one line segment for each $r \in R$, ensured by Constraints (4.38).

$$\sum_{k=1}^{\mathscr{K}-1} \lambda^k_r = 1, \qquad \forall r \in R \tag{4.38}$$

The following constraints describe the relationship between two consecutive breakpoints and the active interval in between.

$$\theta^1_r \leq \lambda^1_r, \qquad \forall r \in R \tag{4.39}$$

$$\theta^k_r \leq \lambda^{k-1}_r + \lambda^k_r, \qquad \forall r \in R, k \in \{2,\ldots,\mathscr{K}-1\} \tag{4.40}$$

$$\theta^{\mathscr{K}}_r \leq \lambda^{\mathscr{K}-1}_r, \qquad \forall r \in R \tag{4.41}$$

**Breakpoints determination**

The approximation error from the outer-inner approximation can be reduced by using more breakpoints in the area where the nonlinear function has higher curvature. However, using too many breakpoints will significantly increase the number of variables and constraints, resulting in a heavy computational burden. In this section, we propose a breakpoint determination method with the aim of locating the fewest breakpoints with a good distribution so that a certain level of approximation accuracy can be guaranteed.

First of all, a maximum acceptable approximation (MAA) error for each interval needs to be specified as the threshold, denoted by $\gamma$. Then, the maximum approximation error between two consecutive breakpoints can be calculated given the equation of the logarithmic function and the approximation functions. Details on how to calculate the maximum approximation error are introduced later in this section. Next, specifying the coordinate of one breakpoint, the location of another breakpoint can be determined by ensuring that the maximum approximation error within the interval formed by these two breakpoints does not exceed the predetermined MAA error threshold $\gamma$. In this case, we can start from the last known breakpoint which is (1, 0) for the logarithmic function $\ln P^r_{AV}$, and then calculate the coordinate of the previous breakpoint. This procedure repeats until the x-coordinate of the newly found breakpoint is smaller than the lower bound of $P^r_{AV}$ for all the groups $r \in R$, which is denoted as $\underline{P}$. $\underline{P}$ is defined as the minimum probability of choosing SAVs among all the groups, under which the demand for SAVs in all the groups will be zero. Namely, under $\underline{P}$, it is pointless to add additional breakpoints.

According to Constraints (4.5), when the value of $n^r P^r_{AV}$ is less than 0.5, the value of $D^r_{AV}$ will be zero. Thus, $\underline{P}$ can take the largest value which satisfies Constraints (4.42).

$$\underline{P} \leq \frac{1}{2n^r}, \qquad \forall r \in R \tag{4.42}$$

When the value of $P^r_{AV}$ is less than $\underline{P}$, the demand for SAVs in group $r \in R$ will reach zero. However, this does not mean that $P^r_{AV}$ cannot have a value less than $\underline{P}$. When the difference between the utility of SAVs and bicycles is sufficiently large, the probability of choosing SAVs may drop to near zero. To ensure the feasibility of the model, a boundary breakpoint must be added. The coordinates of this breakpoint can be specified as $(1/\mathscr{M}, \ln(1/\mathscr{M}))$, where $\mathscr{M}$ is a sufficiently large number.

As shown in Figure 4.4.2, the approximated value lies between the tangent lines (yellow lines) and the secant lines (blue lines), while the real value is the logarithmic function (black line). Thus, the maximum approximation error is the maximum of 1) the maximum distance between the logarithmic function and the secant line, denoted as

$e_1^{\max}$, and 2) the maximum distance between the logarithmic function and the tangent lines, denoted as $e_2^{\max}$. The maximum approximation error takes the maximum value between $e_1^{\max}$ and $e_2^{\max}$ which yields:

$$e^{\max} = \max\{e_1^{\max}, e_2^{\max}\}. \tag{4.43}$$

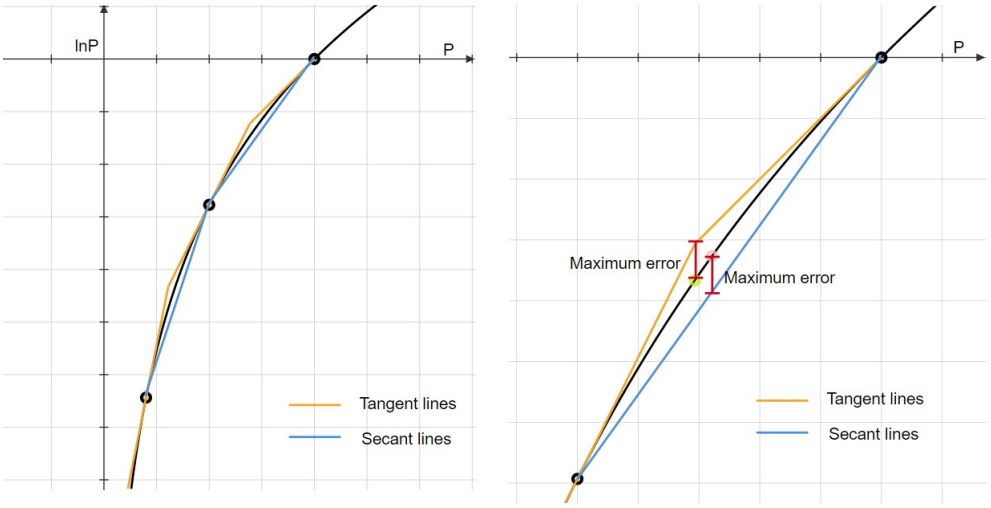



*Figure 4.4.2: Maximum approximation error.*

**Proposition 4.4.1** *The maximum error between the logarithmic function and the secant line $e_1^{\max}$ equals the maximum error between the logarithmic function and the tangent lines $e_2^{\max}$. The maximum error at interval $[u^{k-1}, u^k]$ is:*

$$e^{\max} = \frac{u^{k-1}\ln u^k - u^k \ln u^{k-1}}{u^k - u^{k-1}} - \ln\left(\ln u^k - \ln u^{k-1}\right) + \ln\left(u^k - u^{k-1}\right) - 1. \tag{4.44}$$

**Proof :**
We first calculate the maximum approximation error between the logarithmic function and the secant line.

For $P_{AV}^r \in [u^{k-1}, u^k]$, we define the error at $P_{AV}^r$ between the approximated value and the real logarithmic function as $e$, where

$$e = \ln P_{AV}^r - \left(\ln u^{k-1} + \frac{\ln u^k - \ln u^{k-1}}{u^k - u^{k-1}}\left(P_{AV}^r - u^{k-1}\right)\right). \tag{4.45}$$

To determine which point between these two breakpoints contributes to the maximum error, we have to set the derivative of the error $e$ to 0.

$$\frac{1}{P_{AV}^r} - \frac{\ln u^k - \ln u^{k-1}}{u^k - u^{k-1}} = 0. \tag{4.46}$$

This results in

$$P_{AV}^r = \frac{u^k - u^{k-1}}{\ln u^k - \ln u^{k-1}}. \tag{4.47}$$

At this point, the maximum approximation error occurs. The approximation error equals the difference between the logarithmic function and the secant line at this point.

$$e_1^{\max} = \ln\left(u^k - u^{k-1}\right) - \ln\left(\ln u^k - \ln u^{k-1}\right) + \frac{u^{k-1}\ln u^k - u^k \ln u^{k-1}}{u^k - u^{k-1}} - 1. \tag{4.48}$$

The maximum approximation error between the tangent lines and the logarithmic function occurs at the point where two tangent lines of the consecutive breakpoints intersect. Knowing the coordinates of the two consecutive breakpoints $(u^k, \ln u^k)$ and $(u^{k-1}, \ln u^k)$ with $u^{k-1} < u^k$, the tangent lines at those two breakpoints can be expressed as follows:

$$y^k = \frac{1}{u^k}x + \ln u^k - 1, \tag{4.49}$$

$$y^{k-1} = \frac{1}{u^{k-1}}x + \ln u^{k-1} - 1. \tag{4.50}$$

Combining the two equations, we can calculate the intersection point of the two tangent lines, which is

$$\left(\frac{(\ln u^k - \ln u^{k-1})u^k u^{k-1}}{u^k - u^{k-1}}, \frac{(\ln u^k - \ln u^{k-1})u^{k-1}}{u^k - u^{k-1}} + \ln u^k - 1\right).$$

The maximum approximation error $e_2^{\max}$ is the vertical distance from the intersection point to the logarithmic function:

$$e_2^{\max} = \frac{\left(\ln u^k - \ln u^{k-1}\right)u^{k-1}}{u^k - u^{k-1}} + \ln u^k - 1 - \ln\left(\frac{\left(\ln u^k - \ln u^{k-1}\right)u^k u^{k-1}}{u^k - u^{k-1}}\right). \tag{4.51}$$

This gives:

$$e_2^{\max} = \frac{u^{k-1}\ln u^k - u^k \ln u^{k-1}}{u^k - u^{k-1}} - \ln\left(\ln u^k - \ln u^{k-1}\right) + \ln\left(u^k - u^{k-1}\right) - 1. \quad (4.52)$$

So, we have $e^{\max} = e_1^{\max} = e_2^{\max}$, with the maximum approximation errors occurring at different locations.

□

Setting the last breakpoint equal to $(1,0)$ and fixing the desired maximum error $\gamma$, we can determine the breakpoint before 1 by solving the following formula numerically for $u^{k-1}$.

$$\gamma = \ln\left(u^k - u^{k-1}\right) - \ln\left(\ln u^k - \ln u^{k-1}\right) + \frac{u^{k-1}\ln u^k - u^k \ln u^{k-1}}{u^k - u^{k-1}} - 1 \quad (4.53)$$

with

$$u^k = 1. \quad (4.54)$$

Similarly, $u^{k-2}$ can be obtained by using the found $u^{k-1}$ as input.

### 4.4.2 Linearisation of the floor function

The demand calculation function in Equations (4.5) is non-linear. Therefore, we replace Equations (4.5) by the following constraints:

$$n^r \cdot P_{AV}^r - 0.5 < D_{AV}^r \leq n^r P_{AV}^r + 0.5, \qquad \forall r \in R. \quad (4.55)$$

### 4.4.3 Linearisation of the acceptance rate constraint

When an SAV operator is allowed to reject nonprofitable requests, $\alpha$ is defined as a continuous variable with $\alpha \in [0,1]$. As a result, Constraint (4.6) becomes a non-linear constraint consisting of the product of the continuous variable $\alpha$ and the integer variables $D_{AV}^r$. To linearise this constraint, we introduce additional binary variables $\overline{D}_h$ to discretise the integer term $\sum_{r\in R} D_{AV}^r$, and continuous variables $Y_h \in [0,1]$ to describe the value of the integer term $\sum_{r\in R} S^r$, where $h \in \{0,1,\ldots,\mathscr{H}\}$. Here, $\mathscr{H}$ should be chosen such that these constraints still hold when all demand would use SAVs.

Then, we substitute Constraint (4.6) with the following constraints.

$$\sum_{r\in R} D_{AV}^r = \sum_{h=0}^{\mathscr{H}} 2^h \overline{D}_h \quad (4.56)$$

$$Y_h \leq \alpha, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.57}$$

$$Y_h \leq \overline{D}_h, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.58}$$

$$Y_h \geq \alpha + \overline{D}_h - 1, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.59}$$

$$\sum_{h=0}^{\mathscr{H}} 2^h Y_h = \sum_{r \in R} S^r \tag{4.60}$$

### 4.4.4 Tightening the model by choosing an appropriate value for $\mathscr{M}$

$\mathscr{M}$ used in Constraints (4.21) and (4.26) represents a sufficiently large number. However, using an excessively large $\mathscr{M}$ may lead to a model with a weak relaxation, which can, in turn, slow down the solving process of the mixed-integer programming (MIP) model. Thus, choosing a proper value for $\mathscr{M}$ is beneficial in tightening the proposed model. The main criterion for choosing an appropriate value for $\mathscr{M}$ is to identify the smallest value that is sufficiently large to prevent the cut-off of any feasible solution. The value of $\mathscr{M}$ should be specified for each of the constraints to get a tighter formulation.

We first rewrite Constraints (4.21) as follows with constraint-specific values $\mathscr{M}_r^1$ where $r \in R$.

$$T_{AV}^r \leq A_t^r (t - a^r) + \mathscr{M}_r^1 (1 - Z_t^r), \qquad \forall r \in R, a^r + st^r \leq t \leq b^r \tag{4.61}$$

Given that the longest travel time $T_{AV}^r$ for SAV in group $r \in R$ is inherently less than or equal to the time difference between the latest arrival time $b^r$ and the departure time $a^r$, $\mathscr{M}_r^1$ can take the following value regardless of the values of the binary variables $A_t^r$.

$$\mathscr{M}_r^1 = b^r - a^r, \qquad \forall r \in R \tag{4.62}$$

Then, we rewrite Constraints (4.26) using a constraint-specific value $\mathscr{M}_{ijt_1t_2}^2$ with $t_1, t_2 \in T$ if $t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min}, (i,j) \in L$.

$$t_1 + \sum_{t \in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \sum_{t \in T} X_{i_{t_2} j_t} (t - t_2) + \mathscr{M}_{ijt_1t_2}^2 \left(1 - \sum_{t \in T} X_{i_{t_2} j_t}\right), \qquad \forall (i,j) \in L, \; t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min} \tag{4.63}$$

When the value of $\sum_{t\in T} X_{i_{t_2} j_t}$ is 0, indicating that no vehicles enter the link $(i,j) \in L$ at time instant $t_2 \in T$, Constraints (4.63) become the follows:

$$t_1 + \sum_{t\in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \mathscr{M}^2_{ijt_1t_2}, \qquad \forall (i,j) \in L, t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij}. \quad (4.64)$$

The left-hand side of Constraints (4.64) indicates the time that a vehicle leaves link $(i,j) \in L$ if it enters this link at time instant $t_1 \in T$. Knowing that a maximum travel time for a vehicle traversing link $(i,j) \in L$ is $t^{\max}_{ij}$, the latest time that a vehicle leaves link $(i,j) \in L$ can never exceed its maximum travel time plus the entering time, which gives:

$$t_1 + \sum_{t\in T} X_{i_{t_1} j_t} (t - t_1) \leq t_1 + t^{\max}_{ij}, \qquad \forall t_1 \in T, (i,j) \in L. \quad (4.65)$$

Combining Constraints (4.64) and (4.65) gives the smallest value that $\mathscr{M}^2_{ijt_1t_2}$ can take, which is:

$$\mathscr{M}^2_{ijt_1t_2} = t_1 + t^{\max}_{ij} - t_2, \qquad \forall (i,j) \in L, t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij}. \quad (4.66)$$

To help readers comprehend the model more efficiently, we summarise the complete problem formulation as well as the notations of the sets, parameters, and variables in Appendix 4.A.

## 4.5 Case study of the city of Delft, in The Netherlands

In this section, we present the computational results of the case study of Delft, in the Netherlands to evaluate the effectiveness of the proposed model.

### 4.5.1 Application setting

The proposed model is applied to a quasi-real case study of the city of Delft, in the South Holland province in The Netherlands (Correia & Van Arem, 2016). A simplified road network of Delft is used in this case study which contains 35 nodes and 104 directed links (two-way circulation allowed), displayed in Figure 4.5.1. SAVs are free to drive on the entire network, but only 7 nodes are designated as free parking depots, which are nodes 3, 10, 11, 15, 19, 22, and 27 (identified in red). The parking depots are distributed throughout the city, with three located in the city centre and four located on the outskirts, which facilitates the use of SAV services by residents from all areas of the city. In addition, each road link has either one lane or two lanes, with a capacity of 1600 or 3200, respectively. Vehicles are allowed to travel on these two types of road

links with a maximum travel speed (free-flow speed without congestion) of 50km/h and 70 km/h, respectively, and with a minimum travel speed of 5km/h.

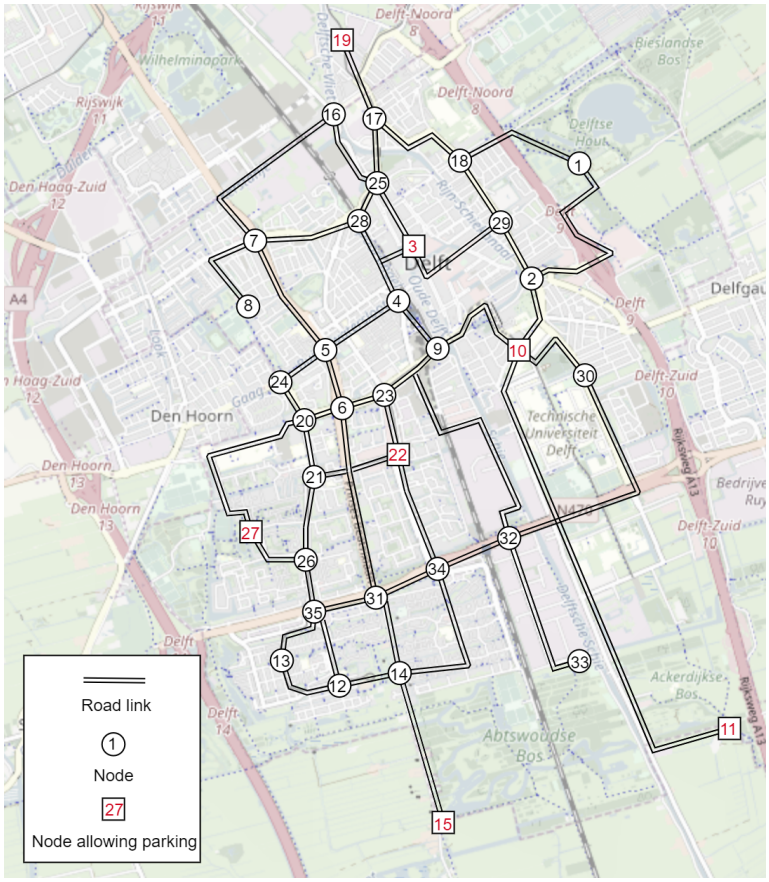



*Figure 4.5.1: Simplified road network of Delft used in the case study.*

The mobility data for the morning peak hour in Delft was obtained using the Dutch mobility dataset (MON 2007/2008). This dataset provides the daily mobility information of a sample of residents, including but not limited to the origin, destination, departure time, arrival time, transport mode, etc. It has been used previously (Correia & Van Arem, 2016), Liang et al. (2020) to study the future mobility system with AVs in urban networks. However, this dataset does not have a large sample of trips for this city if we focus on just one hour. To overcome this limitation and characterise as much as possible the real mobility pattern in the morning peak hour, we filtered out the trips in the database from 7 am to 10 am with travel modes of bicycle, car, and taxi, and then evenly distributed them within one hour with one-third of the amount. In total, 2933 trips are generated, aggregated into 45 groups of trips by the similarity of the trip

information as explained in the model.

The optimisation period contains two parts. One is a one-hour period studied during the morning peak, comprised of 24 time steps of each 2.5 minutes. Besides that, 5 additional time steps are added as a pre-optimisation period. This is needed since we assume that all the SAVs depart from parking depots in the morning to serve the trips. For SAVs to arrive at the requested origin on time, additional time steps are required as slack in the optimisation period. Therefore, the optimisation period contains a total of 29 time steps.

The parameters used in this case study related to the network setting, the demand, and the SAV operating system are summarised and explained in Table 4.5.1.

*Table 4.5.1: Parameter summary*

| Parameter | Description and reference values |
|---|---|
| $t_{ij}^{\max}$ | Maximum travel time by SAVs which is computed by dividing the length of the road link by the minimum travel speed of 5km/h. |
| $t_{ij}^{\min}$ | Minimum travel time by SAVs which is calculated by dividing the length of the road link by the related maximum travel speed (50km/h or 70km/h). Note that the minimum travel time on each road link has a minimum value of 1 time step (2.5 minutes in this case study) due to the time-space network. It imposes that no vehicles can travel with a travel time of zero. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of each road link which is calculated using Equation (4.23). |
| $sd^r$, $st^r$ | Shortest travel distance/time which is calculated using the shortest path algorithm assuming SAVs can travel with free-flow speed. |
| $T_B^r$ | Travel time of bicycles which is calculated by dividing the length of the shortest path by the average speed of the Dutch on a pedal bicycle, 12.4 km/h (BicycleDutch, 2018). |
| $\beta_0$ | Parameter used in the logit model with a value of 0.1. |
| $\beta_{AV}^r$ | Travellers' VOTT for using an AV with high income, middle income, and low income equal to 6.6, 4.6, and 3.8 euro/hour, respectively (Kolarova et al., 2019). |
| $\beta_B^r$ | Travellers' VOTT for using a bicycle with high income, middle income, and low income equal to 24.9, 17.3, and 14.1 euro/hour, respectively (Kolarova et al., 2019). |
| $p^0$, $p$ | Initial base fare and price per km which are set to 3 euros and 1.28 euros/km, respectively, according to the price rate of Uber in Delft, in The Netherlands (Uber, 2023). |
| $co$ | Operational cost of SAVs which is set to 0.32 euro/km (calculated according to the methodology proposed by Bösch et al. (2018). |
| $cf$ | Depreciation cost of SAVs which is set to 1.2 euro/vehicle/hour (Fan et al., 2022). |
| $cd$ | Delay penalty which is set to 0.2 euro/min (Liang et al., 2020). |
| $\alpha$ | Service rate which is set to 1 when the SAV operator has to serve all the trips. |
| $a, b$ | Parameters in the BPR function which are set to 2 and 4, respectively (Van Essen & Correia, 2019). |

### 4.5.2 Breakpoint generation

Before solving the reformulated MILP model, a set of breakpoints was generated with a pre-specified MAA error. This error represents the maximum acceptable difference between the approximation value and the true value of the non-linear terms in Constraints (4.31) and (4.32). Therefore, the smaller this value is, the more precise the approximation will be; however, the greater the number of breakpoints that will be generated.

Figure 4.5.2 shows the relationship between the number of generated breakpoints and the value of the MAA error. We can observe a clear trend where the number of generated breakpoints increases dramatically with the decrease in the value of the MAA error, especially when the approximation error is less than 0.05. On the one hand, more breakpoints lead to higher accuracy, but on the other hand, they lead to greater computational time. To balance these two factors, we need an MAA error that can ensure a good quality of the optimisation results within an acceptable computational time. To find a proper value for this case study, we first tested the model in the base scenario with three different values for the MAA error, which are 0.05, 0.01, and 0.005, yielding 11, 23, and 31 breakpoints, respectively.

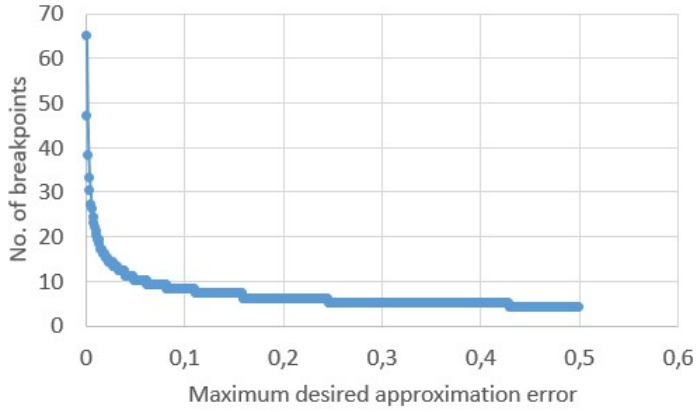



*Figure 4.5.2: Relationship between MAA error and the number of generated breakpoints .*

We implemented the reformulated MILP model in Python 3.7 and then solved it using Gurobi optimiser version 10.0.0 on an on an Intel(R) Xeon(R) W-2123 CPU @3.60 GHz, and 32.00 GB RAM computer. Table 4.5.2 shows the optimisation results with the three mentioned MAA errors. Here, we used the objective function value, the optimal total fleet size, the total demand for SAVs, and the computational time as indicators to compare the performance for these three cases. The objective function

value is the maximised profit for an SAV operator. The optimal fleet size and the total demand for SAVs are selected as the indicators because fleet sizing is one of the most important planning decisions for an SAV operator, and the demand for SAVs directly impacts the fleet sizing decision. In addition to this reason, the demand for SAVs is one of the attributes most affected by the approximation error.

*Table 4.5.2: optimisation results with different MAA errors*

| MAA error | 0.005 | 0.01 | | 0.05 | |
|---|---|---|---|---|---|
| | Value | Value | Relative change | Value | Relative change |
| Number of breakpoints | 31 | 23 | -28.57% | 11 | -64.29% |
| Objective function value | 11128.07 | 11143.30 | +0.14% | 11414.46 | +2.57% |
| Optimal fleet size of SAVs | 890 | 891 | +0.11% | 914 | +2.70% |
| Demand for SAVs | 1269 | 1271 | +0.16% | 1304 | +2.76% |
| Computational time | 3126s | 2236s | -28.47% | 1921s | -38.55% |

We first ran the model with MAA error of 0.005, then used the corresponding optimisation results as the benchmark to compare with other cases in which the MAA error is 0.01 and 0.05. The relative changes in the values of the indicators are computed. Looking at the optimisation results with MAA values of 0.005 and 0.01 in Table 4.5.2, we observe very small differences in the objective function values (0.14% relative difference), the values of the optimal fleet sizes (1 unit difference), and the values of the demand for SAVs (2 units difference), indicating that using 23 breakpoints has already achieved a good approximation accuracy. Adding more breakpoints does not bring a significant improvement to the optimisation outcomes.

In all, the MAA error of 0.01 was used throughout the experiments, which yields 23 breakpoints.

### 4.5.3 Optimisation results

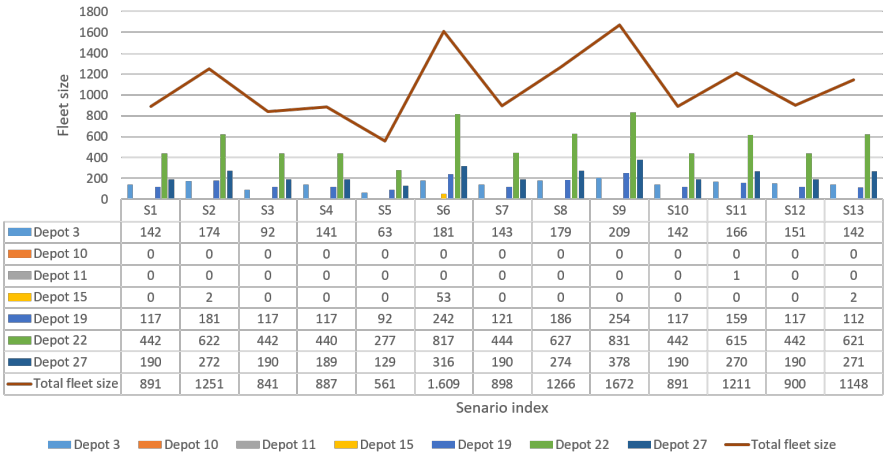
The model was tested first in a base scenario with parameters given in Section 4.5.1. Then, we conducted a sensitivity analysis to the following parameters: SAVs price rate, unit operational cost, delay penalty, parameter $\beta_0$, and the combination of them in 6 scenarios. We also investigate the impact of congestion by evaluating non-congested scenarios as a comparison of the existing scenarios. As previously mentioned, this chapter explores two accept/reject mechanisms. Scenarios 1 to 9 assume that the SAV operator must accept all requests, while scenarios 10 to 13 assume that the SAV operator may reject the nonprofitable trips. A sensitivity analysis of SAV price rate and

parameter $\beta_1$ is carried out to see how travellers' satisfaction with the service quality level influences the managerial decisions of the SAV operator under different pricing policies.

Table 4.5.3 shows the descriptions and parameter settings for all scenarios. The optimal fleet size distribution can be found in Figure 4.5.3 and key performance indicators can be found in Table 4.5.4.

*Table 4.5.3: Scenario description*

| Scenario description | $p^0$ (euro) | $p$ (euro/km) | $co$ (euro/km) | $\alpha$ | $cd$ (euro/min) | $\beta_0$ (utility/euro) | $\beta_1$ |
|---|---|---|---|---|---|---|---|
| S1 Base scenario | 3 | 1.28 | 0.32 | 1 | 0.2 | 0.1 | - |
| S2 Lower price | 1.5 | 0.64 | 0.32 | 1 | 0.2 | 0.1 | - |
| S3 Lower operational cost | 3 | 1.28 | 0.1 | 1 | 0.2 | 0.1 | - |
| S4 No delay penalty | 3 | 1.28 | 0.32 | 1 | 0 | 0.1 | - |
| S5 Higher $\beta_0$ | 3 | 1.28 | 0.32 | 1 | 0.2 | 0.5 | - |
| S6 Higher $\beta_0$ with lower price | 1.5 | 0.64 | 0.32 | 1 | 0.2 | 0.5 | - |
| S7 Base scenario without congestion | 3 | 1.28 | 0.32 | 1 | 0.2 | 0.1 | - |
| S8 Lower price without congestion | 1.5 | 0.64 | 0.32 | 1 | 0.2 | 0.1 | - |
| S9 Higher $\beta_0$ with lower price without congestion | 1.5 | 0.64 | 0.32 | 1 | 0.2 | 0.5 | - |
| S10 Base scenario with rejection | 3 | 1.28 | 0.32 | - | 0.2 | 0.5 | 1 |
| S11 Lower price with rejection | 1.5 | 0.64 | 0.32 | - | 0.2 | 0.5 | 1 |
| S12 Base scenario with rejection and lower $\beta_1$ | 3 | 1.28 | 0.32 | - | 0.2 | 0.5 | 0.1 |
| S13 Lower price with rejection and lower $\beta_1$ | 1.5 | 0.64 | 0.32 | - | 0.2 | 0.5 | 0.1 |




| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Depot 3 | 142 | 174 | 92 | 141 | 63 | 181 | 143 | 179 | 209 | 142 | 166 | 151 | 142 |
| Depot 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Depot 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Depot 15 | 0 | 2 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| Depot 19 | 117 | 181 | 117 | 117 | 92 | 242 | 121 | 186 | 254 | 117 | 159 | 117 | 112 |
| Depot 22 | 442 | 622 | 442 | 440 | 277 | 817 | 444 | 627 | 831 | 442 | 615 | 442 | 621 |
| Depot 27 | 190 | 272 | 190 | 189 | 129 | 316 | 190 | 274 | 378 | 190 | 270 | 190 | 271 |
| Total fleet size | 891 | 1251 | 841 | 887 | 561 | 1.609 | 898 | 1266 | 1672 | 891 | 1211 | 900 | 1148 |

*Figure 4.5.3: Fleet size and initial distribution of SAVs at the beginning of a day in all the scenarios.*

*Table 4.5.4: Optimisation results for all the scenarios*

| Scenario | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total profit (euro) | 11143.31 | 3389.07 | 13904.24 | 11695.33 | 7530.06 | 3653.06 | 11924.48 | 4905.77 | 6437.42 | 11143.31 | 3463.92 | 11152.51 | 3505.19 |
| Total revenue (euro) | 16663.61 | 11729.52 | 16672.16 | 16601.24 | 10846.61 | 15113.09 | 16836.59 | 11874.50 | 15667.01 | 16663.61 | 11159.4 | 16599.62 | 10691.84 |
| Average price per trip (euro) | 13.11 | 6.84 | 13.11 | 13.10 | 14.23 | 7.07 | 13.12 | 6.84 | 7.02 | 13.11 | 6.92 | 13.15 | 6.81 |
| Total depreciation cost (euro) | 1069.20 | 1501.20 | 1009.20 | 1064.40 | 673.20 | 1930.80 | 1077.60 | 1519.20 | 2006.40 | 1069.20 | 1453.2 | 1080.0 | 1377.6 |
| Total operational cost (euro) | 3962.10 | 5765.25 | 1285.72 | 3841.51 | 2559.35 | 7510.24 | 3834.51 | 5449.54 | 7223.18 | 3962.10 | 5485.27 | 3934.11 | 5245.55 |
| Total delay penalty cost (euro) | 489 | 1074 | 473 | 0 | 84 | 2019 | 0 | 0 | 0 | 489 | 757.0 | 433 | 563.5 |
| Total demand for SAVs | 1271 | 1715 | 1272 | 1267 | 762 | 2138 | 1283 | 1737 | 2232 | 1271 | 1690 | 1273 | 1714.0 |
| SAV demand share | 43.33% | 58.47% | 43.37% | 43.20% | 25.98% | 72.89% | 43.74% | 59.22% | 76.10% | 43.33% | 57.62% | 43.4% | 58.44% |
| Total satisfied trips for SAVs | 1271 | 1715 | 1272 | 1267 | 762 | 2138 | 1283 | 1737 | 2232 | 1271 | 1613 | 1262 | 1570 |
| Percentage of satisfied demand | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 95.44% | 99.14% | 91.6% |
| Average delay per trip (minute) | 1.93 | 3.13 | 1.85 | 3.25 | 0.55 | 4.73 | 0 | 0 | 0 | 1.93 | 2.35 | 1.68 | 1.8 |
| SAVs total travel distance (km) | 12381.57 | 18016.39 | 12857.21 | 12004.73 | 7997.96 | 23469.50 | 11982.85 | 17029.80 | 22572.45 | 12381.57 | 17141.48 | 12294.10 | 16392.34 |
| SAVs total deliver distance (km) | 10552.54 | 15554.53 | 10807.14 | 10324.48 | 6843.03 | 20266.68 | 10146.56 | 14482.82 | 19248.45 | 10552.54 | 14907.98 | 10523.65 | 14041.77 |
| SAVs total relocate distance (km) | 1829.03 | 2461.86 | 2050.06 | 1680.25 | 1154.92 | 3202.82 | 1836.29 | 2546.98 | 3324 | 1829.03 | 2233.49 | 1770.46 | 2350.57 |
| SAVs total delivery time (hour) | 339.5 | 512.08 | 338.29 | 366.25 | 201.75 | 715.04 | 301.5 | 427.33 | 567.21 | 339.5 | 467.96 | 334.08 | 435.54 |
| Average delivery time per trip (minute) | 16.03 | 17.93 | 15.95 | 17.35 | 15.88 | 20.08 | 14.1 | 14.75 | 15.25 | 16.03 | 17.4 | 15.88 | 16.65 |
| Computational time (s) | 1986s | 7846s | 2398s | 91545s | 634s | 86400s | 337s | 386s | 388s | 9517s | 60758s | 6145s | 11510s |
| MIP Gap | 0 | 0 | 0 | 0 | 0 | 0.57% | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

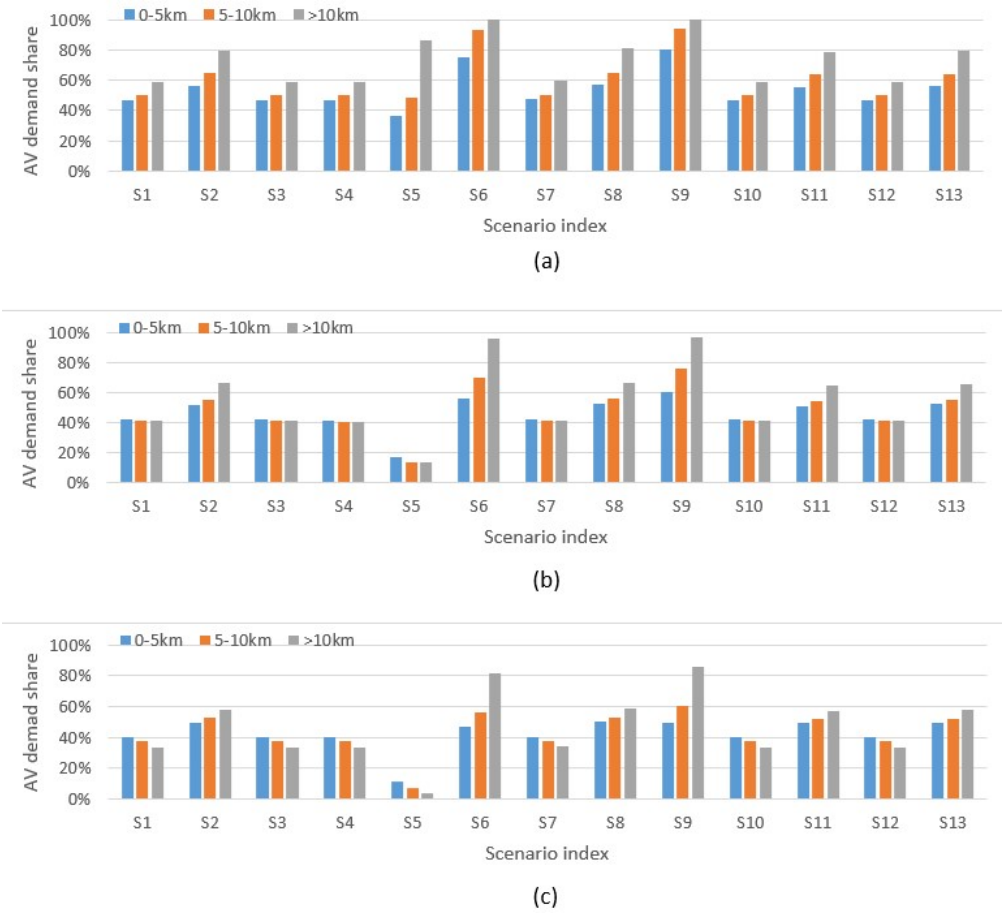



*Figure 4.5.4: SAV demand share for different user classes with (a) high VOTT, (b) middle VOTT, and (c) low VOTT.*

### Base scenario

As can be seen in Figure 4.5.3, almost all the SAVs are distributed at parking depots 3, 19, 22, and 27 at the beginning of the day, as these depots are either close to residential areas or the train station in Delft where the commuting demand is high during the morning peak hour. Depots 10, 11, and 15 have hardly any SAVs as these depots are either located on the outskirts of the city or near the campus area which is usually the destination of commuting in the morning. The distribution of the fleet at the beginning of the day is highly influenced by the geographical distribution of the population, the distribution of land use, and the travel patterns of residents.

In Table 4.5.4, 43.33% of the travellers choose to use SAV services. However,

travellers from different income classes behave differently facing trips with different lengths, as shown in Table 4.5.5. We classified trips into three groups in terms of their lengths: less than 5 kilometres, between 5 and 10 kilometres, and more than 10 kilometres. Then, we calculated the demand share of SAV services for travellers in different classes (with high VOTT, middle VOTT, and low VOTT), and their corresponding cost structures (price, travel time-related cost for using an SAV, and a bicycle).

Results indicate that travellers with a high VOTT are more sensitive to variations in trip length compared with the other classes. When the trip length is longer than 10 kilometres, 59.22% of travellers with a high VOTT use SAVs rather than cycling because the increase in time-related costs of cycling is significant for them. When the length of the trip is short (less than 5 kilometres), 46.9% of travellers with a high VOTT choose SAVs, meaning that using bicycles can slightly save them some costs. But the difference between using these two modes is not big. For trips between 5km and 10km, half of the people choose SAVs as the utilities for using these two modes are the same. It makes little difference which mode they choose. Note that travellers with a middle VOTT and a low VOTT always prefer bicycles to SAVs as the price for using SAVs is high. In addition, travellers with a middle VOTT are insensitive to the changes in trip length. For them, the cost difference between these two modes does not change significantly with the increase in trip length.

*Table 4.5.5: Optimisation results under the base scenario*

| User class | | 0-5km | 5-10km | $\geq$ 10km |
|---|---|---|---|---|
| High VOTT | SAV demand share | 46.90% | 50.23% | 59.22% |
| | Average price per trip | 8.04 | 12.32 | 20.81 |
| | Average travel time-related cost per trip for using an SAV | 0.99 | 1.79 | 2.70 |
| | Average travel time-related cost per trip for using a bicycle | 7.53 | 13.82 | 27.03 |
| Middle VOTT | SAV demand share | 41.59% | 40.93% | 40.78% |
| | Average price per trip | 8.04 | 12.32 | 20.81 |
| | Average travel time-related cost per trip for using an SAV | 0.76 | 1.36 | 1.98 |
| | Average travel time-related cost per trip for using a bicycle | 5.23 | 9.63 | 18.78 |
| Low VOTT | SAV demand share | 40.00% | 37.33% | 33.33% |
| | Average price per trip | 8.04 | 12.32 | 20.81 |
| | Average travel time-related cost per trip for using an SAV | 0.62 | 1.07 | 1.75 |
| | Average travel time-related cost per trip for using a bicycle | 4.27 | 7.85 | 15.31 |

**Sensitivity analysis on price rate**

The price rate has a great impact on travellers' behaviour, which in turn affects the total demand for SAV services and fleet sizing decisions. In Table 4.5.4, one can see that the demand for SAVs increased from 1271 (in S1 Base scenario) to 1715 (in S2 Lower price) when the price rate reduces to half of what it was in the base scenario S1.

To satisfy the increased demand, the SAV operator has to deploy a larger fleet size. At the same time, the congestion level increased as more vehicles circulated on the road network and competed for the shortest paths. It can be seen from Table 4.5.4 that the average delay time per trip increased from 1.93 (in S1 Base scenario) to 3.13 minutes (in S2 Lower price) and the average delivery time per trip increased from 16.03 minutes (in S1 Base scenario) to 17.93 minutes (in S2 Lower price). Thus, more delay penalty was generated which reduced the total profits. Even though more travellers choose to use SAVs, the total profit is still lower than the one in S1 Base scenario because of the lower revenue, higher depreciation costs, higher operational costs and higher delay penalty.

When the price rate for using an SAV is reduced, the willingness of travellers from different income classes to use the SAV service increases compared with S1 Base scenario, as can be seen in Figure 4.5.4. For travellers with high VOTT (in Figure 4.5.4a), SAVs are preferable to bicycles regardless of trip length. For travellers with middle VOTT whose trips are less than 10km in length, the preference between the two modes is not obvious. Only when the trip length exceeds 10km, do travellers prefer to use SAVs over the bicycle mode. A similar trend is observed for travellers with a lower VOTT. The longer the trip is, the more likely a traveller with a low VOTT will choose SAVs.

**Sensitivity analysis on operational costs**

As depicted in Figure 4.5.3, the total fleet size in S3 Lower operational cost is 50 vehicles smaller than the one in S1 Base scenario. Looking at the optimisation results displayed in Table 4.5.4, we found that the total profit increases from 11143.31 euros (in S1 Base scenario) to 13904.24 euros (in S3 Lower operational cost). The change in demand for SAVs in S1 and S3 is negligible, and the total relocation distance increases from 1829.03 km (in S1 Base scenario) to 2050.06 km (in S3 Lower operational cost). This indicates that the SAV operator can save money by deploying a smaller fleet and allowing SAVs to relocate more frequently at a lower operational cost.

In addition, the total delay penalty cost decreases from 489 euros (in S1 Base scenario) to 473 euros (in S3 Lower operational cost), while the total delivery distance of the SAVs increases from 10552.54 km (in S1 Base scenario) to 10807.14 km (in S3 Lower operational cost), meaning that SAVs detour more to avoid traffic congestion to deliver clients as soon as possible, as well as to pick up more clients.

Looking at Figure 4.5.4, we barely notice any difference between the SAV shares for different user classes and trip length. Thus, we conclude that lowering the operational cost of the SAV fleet does not have a significant influence on the demand

structure.

All in all, SAV operators earn more profits through operational cost savings, less delay penalty, and fewer depreciation costs of the fleet, even though SAVs detour and relocate more.

**Sensitivity analysis on delay penalty**

The SAV operator earns greater profits when there is no delay penalty for the late drop-off of clients in S4. However, the attractiveness of the SAV service drops slightly, which is reflected in the reduced demand for SAVs, which can be seen in Table 4.5.4. Furthermore, the reduced amount of fleet size is consistent with the decreased demand for SAVs.

To compare the trip delay information between S4 (no delay penalty) and S1 (base scenario), we have plotted the delay distributions, along with the mean and the 90th percentile values for the delay in Figure 4.5.5. The results indicate that 90% of trips in S1 experience a delay within 5 minutes, whereas 90% of trips in S4 have a delay within 7.5 minutes. Additionally, the average delay per trip is 1.93 minutes in S1 and 3.25 minutes in S4. These findings indicate that when there is no penalty for late deliveries, the actual delivery time becomes longer compared to the base scenario S1, resulting in increased delay, which can also be seen from the increased average delivery time per trip and the increased average delay per trip in S4 in Table 4.5.4. However, it is crucial to consider the perspective of passengers using the SAV service in an inter-modal fashion, who require a certain level of reliability in their arrival time, because of the need to coordinate with other transportation modes. In scenarios where an SAV cannot guarantee arrival before a traveller's acceptable latest arrival time, the trip may be rejected, and the traveller may opt for an alternative mode of transportation. From this perspective, implementing a delay penalty for the SAVs could encourage more reliable and timely deliveries, addressing passengers' concerns and potentially reducing trip rejections. The value of the delay penalty will influence the demand for SAVs, and consequently, impact overall profitability. Investigating the optimal delay penalty value remains an area for future research.

The value of the delay penalty will influence the demand for Shared Autonomous Vehicles (SAVs), and consequently, impact overall profitability. Investigating the optimal delay penalty value and the spatial impact of the rejection rate remains an area for future research."

Overall, the delay penalty does not have a significant impact on fleet sizing decisions and travellers' behaviour. Travellers with higher VOTT care more for late arrival than those with a relatively lower VOTT.

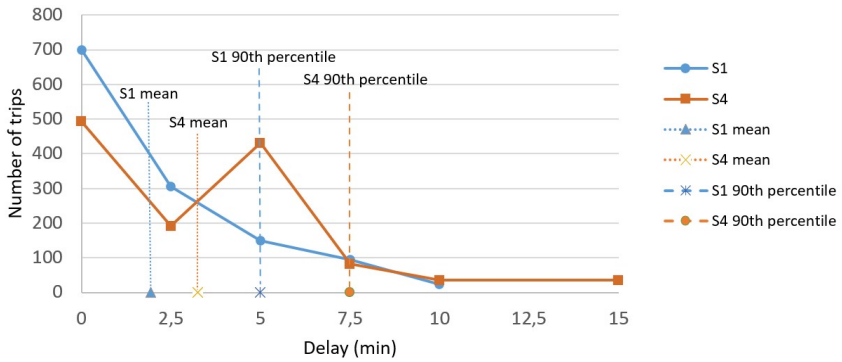



*Figure 4.5.5: Delay distribution in S1 (base scenario) and S4 (no delay penalty).*

### Sensitivity analysis on $\beta_0$

The parameter $\beta_0$ indicates the level of sensitivity that travellers exhibit towards the changes in monetary costs. In this section, we tested a higher value of $\beta_0$ which is 0.5 utility/euro in two scenarios S5 and S6, based on S1 Base scenario and S2 Lower price. We tested these two scenarios with different price rates because the congestion level is different in both, which allows for observing the impact of network congestion levels on the optimisation results.

Looking at Figure 4.5.3, we found a big difference in the optimal fleet size in S5 Higher $\beta_0$ and S6 Higher $\beta_0$ with lower price, compared with S1 Base scenario and S2 Lower price. The fleet size differences mainly come from the decreased/increased demand in these two scenarios, which are 762 and 2138, as can be found in Table 4.5.4. When looking into the details of SAV demand share in S5 Higher $\beta_0$ and S6 Higher $\beta_0$ with lower price, we notice that more travellers tend to choose the mode with the least generalised costs, resulting in a greater difference between the demand share for SAVs and bicycles. Note that the variation of the SAV demand share for all the user classes in all types of trips in S5 and S6 share a similar trend as in S1 and S2, indicating that a higher value of $\beta_0$ can only bring a larger degree of variation to the demand share, but it cannot completely alter travellers' preference towards the modes.

A value of $\beta_0$ that is precisely estimated for the application city can enhance the realism of the model and lead to a more accurate fleet sizing decision. Note that the model was solved in S6 Higher $\beta_0$ with a lower price with a MIP gap of 0.57% with a time limit of 24h because the congestion level is high due to the increased demand for SAVs.

**Impact of traffic congestion**

To test the impact of traffic congestion on the strategic and operational decisions, we removed traffic congestion from S1 Base scenario, S2 Lower price and S6 Higher $\beta_0$ with lower price, by assuming all vehicles can travel at free-flow speed. These three scenarios are selected as references because they exhibit gradually increased congestion levels. Removing congestion in these three scenarios gives us three new scenarios named S7, S8, and S9.

In Figure 4.5.3, it can be seen that the optimal fleet size increases from 891 (in S1 Base scenario) to 898 (in S7 Base scenario without congestion), from 1251 (in S2 Lower price) to 1266 (in S8 Lower price without congestion), from 1609 (in S6 Higher $\beta_0$ with lower price) to 1670 (in S9 Higher $\beta_0$ with a lower price without congestion). It turns out that congestion has a significant impact on fleet sizing decisions, which should be taken into consideration when solving the fleet management problem.

Without congestion, all trips can be delivered to the desired destinations in the shortest travel time and travel distance, as can be seen in Table 4.5.4. The average delay per trip and the total delay penalty in S7 Base scenario without congestion, S8 Lower price without congestion and S9 Higher $\beta_0$ with lower price without congestion are 0. Consequently, the demand for SAVs increased from 1271 (in S1 Base scenario) to 1283 (in S7 Base scenario without congestion), from 1715 (in S2 Lower price) to 1737 (in S8 Lower price without congestion), and from 2138 (in S6 Higher $\beta_0$ with lower price) to 2232 (in S9 Higher $\beta_0$ with lower price without congestion) because travellers are more willing to take SAVs if the travel time is lower. However, despite an increase in demand, the total travel distance and the total delivery time of SAVs decrease correspondingly, indicating that SAVs no longer need to take longer detours to avoid the competition for the shortest paths, which reduces operational costs significantly.

**Comparison of the two accept/reject mechanisms**

In terms of the accept/reject mechanism, from S10 to S13, the SAV operator can reject nonprofitable trips. $\alpha$ is defined as a continuous variable that represents the trip service rate. However, the rejection rate will have an impact on travellers' satisfaction with the SAV service since $\alpha$ is included in the utility calculation. Two pricing rates are tested. S10 and S12 share the same price setting as S1 Base scenario. S11 and S13 share the same price setting as S2 Lower price. Besides, travellers may have different sensitivities to the rejection rate, which is reflected in parameter $\beta_1$. A lower value of parameter $\beta_1$ is tested in S12 and S13 meaning that travellers can have a lower sensitivity towards the rejection rate.

First, we shall have a look at the trip service rate in different scenarios when the SAV operator is allowed to reject non-profitable trips. Looking at the optimisation results in Table 4.5.4, we noticed that S1 Base scenario and S10 Base scenario with rejection yield the same service rate, indicating that the SAV operator did not reject any requests to maintain a high level of service quality despite having the option to decline nonprofitable requests. In S10 Base scenario, $\beta_1$ equals 1, and travellers are sensitive to the change in rejection rate. Thus, with this price setting, rejecting trips can decrease demand for SAVs even for profitable requests. However, in S11, when the price rate of using SAVs is lower than in S10 Base scenario, the trip satisfaction rate dropped to 95.44%. This indicates that the SAV operator is willing to accept the loss of revenue caused by decreased travellers' satisfaction and reduced demand in order to save costs by rejecting non-profitable trips. As can be observed in Table 4.5.4, the SAV operator earned more profits in S11 Lower price with rejection compared with S2 Lower price while satisfying fewer trips. The cost saving comes from less operational cost, less depreciation cost, and less delay penalty.

When travellers are less sensitive to the service quality level, the trip service rate decreased from 100% (in S10 Base scenario with rejection) to 99.14% (in S12 Base scenario and lower $\beta_1$) and from 95.44% (in S11 Lower price with rejection) to 91.6% (in S13 Lower price with rejection and lower $\beta_1$). It indicates that the SAV operator can increase the profit by rejecting more nonprofitable trips, even if it brings negative impacts on travellers' satisfaction. Although this resulted in a decrease in revenue due to a lower number of satisfied trips, the SAV operator can save on operational costs and reduce delay penalties leading to a higher overall profit.

Rejecting some trips mitigates traffic congestion on the network. As shown in Table 4.5.4, the average delay per trip decreases from 1.93 minutes (in S1 Base scenario and S10 Base scenario with rejection) to 1.68 minutes (in S12 Base scenario with rejection and lower $\beta_1$), and the average delivery time per trip decreases from 16.03 minutes (in S1 Base scenario and S10 Base scenario with rejection) to 15.88 minutes (in S12 Base scenario with rejection and lower $\beta_1$). The same trend can be found when the price rate of using SAV services is low. The average delay per trip decreases from 3.13 minutes (in S2 Lower price) to 2.35 minutes (in S11 Lower price with rejection), then to 1.8 minutes (in S13 Lower price with rejection and lower $\beta_1$), and the average delivery time per trip decreases from 17.93 minutes (in S2 Lower price) to 17.4 minutes (in S11 Lower price with rejection), then to 16.65 minutes (in S13 Lower price with rejection and lower $\beta_1$).

In terms of the fleet sizing decisions, we can conclude that these two accept/reject mechanisms do not have a significant impact on the fleet size decisions when the price rate of using SAVs is high. As can be seen in Table 4.5.4, the total SAV fleet size in

S10 Base scenario with rejection is the same as that in S1 Base scenario, while the total SAV fleet size in S12 Base scenario with rejection and lower $\beta_1$ is slightly higher than that in S1 Base scenario and S10 Base scenario with rejection. This indicates that using a bit more vehicles in S12 can save the relocation distance and further release the congestion effect caused by the relocation of SAVs. This part of the savings is greater than the increased depreciation costs of the total fleet which makes it the optimal strategy in S12. However, when the price rate is low, we observe that the fleet size is sensitive to the accept/reject mechanism and parameter $\beta_1$. When travellers have a low sensitivity to the service quality level (rejection rate), the SAV operator tends to reject more nonprofitable trips to gain more profits. Thus, a smaller fleet can be deployed as the number of served trips decreases. Our future research will involve further exploration of the spatial impacts of the rejection rate.

## 4.6 Scaling analysis: model performance with various network sizes and demand

Scalability denotes the capability of the proposed methodology to manage an expanding workload, including accommodating larger network sizes and rising demands. Investigating the scalability of our proposed model holds significance due to its nature as a single-level mixed integer programming model that integrates endogenous demand, congestion and accept/reject mechanism, and that is solved using an exact method. Thus, within this section, we present the computational tests conducted to evaluate the performance of the proposed model under various network sizes and demand profiles. These experiments were executed on a desktop computer with an Intel(R) Xeon(R) W-2123 CPU @3.60 GHz, and 32.00 GB RAM. The implementation of the model was accomplished using Python 3.7, and the MILP solver Gurobi 10.0.0 was utilised to solve the optimisation problems.

To evaluate the model's scalability towards the network sizes, we generated three grid networks, each with varying sizes: 16 nodes and 48 directed links, 64 nodes and 224 directed links, and 144 nodes and 528 directed links, as illustrated in Figure 4.6.1. All the links are two-way circulation allowed. For each network, we distributed 4, 8, and 12 parking depots, respectively. The road links in these networks have an equal length of 2 kilometres and an equal capacity of 3200 vehicles/hour. In our testing, we set a time step of 2.5 minutes with the shortest travel time per link at 2.5 minutes (1 time step) and the longest travel time per link at 10 minutes (4 time steps).

To investigate the impact of demand variations, we conducted tests with different demand profiles. The total number of trips and the number of groups of trips were

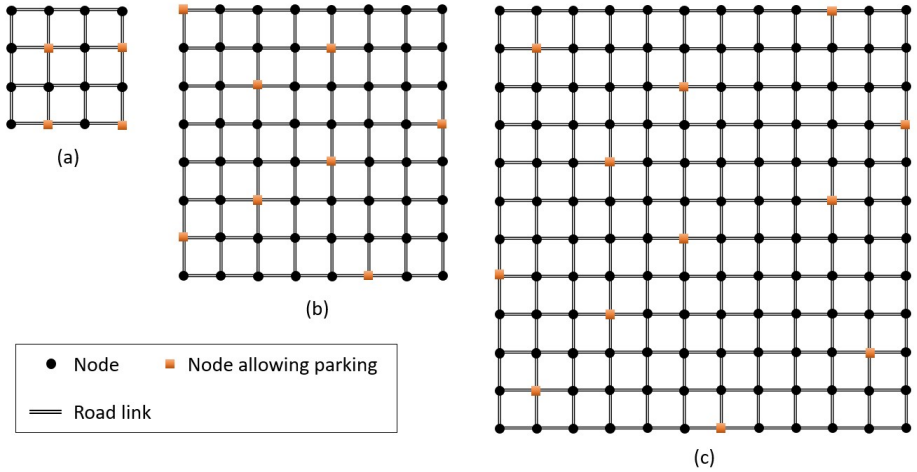



*Figure 4.6.1: Illustration of grid networks: (a) small, (b) medium, and (c) large.*

modified to simulate the varying passenger demands in different scenarios. The configuration of the tests can be found in Table 4.6.1. Trip details such as origin, destination, and departure time were randomly generated to create realistic scenarios. The shortest travel time and the shortest travel distance were calculated using the Dijkstra shortest path algorithm based on the known origin and destination. The latest arrival time for each trip was calculated by doubling the shortest travel time and adding it to the departure time. The optimisation period contains 29 time steps, the same as the case study of the city of Delft in the Netherlands. Furthermore, the remaining parameters remain consistent with the Delft case study and are provided in Table 4.5.1 for reference.

The computation times for each of the tests are summarised in Table 4.6.1. In the first 9 instances, we increase the number of trips and the number of groups of trips for each network setting. Notably, there is a clear tendency for increased computation time with a higher number of trips and more groups. Then, we compare the instances with the same number of trips and groups, but with increased network size from small (N16_L48_P4) to medium (N64_L224_P8) and then to large (N144_L528_P12). For these increased network sizes, we observe a consistent trend of increasing computation time. To test the computational limits of the proposed model using the current computer, we further intensified the congestion level by enlarging the number of groups and the total trips in the large grid network (N144_L528_P12). As depicted in the last 6 instances of Table 4.6.1, this led to a notable increase in computation time. Solving the final instance with 12000 trips and 180 groups of trips within the large grid network took more than 24 hours in computational time. It is important to note that the performance of the proposed model may vary when executed on different computers

*Table 4.6.1: Configurations and computational results (Note: 'N' represents the number of nodes; 'L' represents the number of links; 'P' represents the number of parking depots; 'R' represents the number of trips; 'G' represents the number of group of trips).*

| Configuration | N | L | P | R | G | Computational time | MIP Gap |
|---|---|---|---|---|---|---|---|
| N16_L48_P4_R1000_G30 | 16 | 48 | 4 | 1000 | 30 | 15s | 0 |
| N16_L48_P4_R2000_G60 | 16 | 48 | 4 | 2000 | 60 | 19s | 0 |
| N16_L48_P4_R3000_G90 | 16 | 48 | 4 | 3000 | 90 | 60s | 0 |
| N64_L224_P8_R1000_G30 | 64 | 224 | 8 | 1000 | 30 | $1112\text{s} \approx 0.31\text{h}$ | 0 |
| N64_L224_P8_R2000_G60 | 64 | 224 | 8 | 2000 | 60 | $1270\text{s} \approx 0.35\text{h}$ | 0 |
| N64_L224_P8_R3000_G90 | 64 | 224 | 8 | 3000 | 90 | $7142\text{s} \approx 1.98\text{h}$ | 0 |
| N144_L528_P12_R1000_G30 | 144 | 528 | 12 | 1000 | 30 | $1984\text{s} \approx 0.55\text{h}$ | 0 |
| N144_L528_P12_R2000_G60 | 144 | 528 | 12 | 2000 | 60 | $4192\text{s} \approx 1.16\text{h}$ | 0 |
| N144_L528_P12_R3000_G90 | 144 | 528 | 12 | 3000 | 90 | $9874\text{s} \approx 2.74\text{h}$ | 0 |
| N144_L528_P12_R6000_G90 | 144 | 528 | 12 | 6000 | 90 | $10637\text{s} \approx 2.95\text{h}$ | 0 |
| N144_L528_P12_R9000_G90 | 144 | 528 | 12 | 9000 | 90 | $21925\text{s} \approx 6.09\text{h}$ | 0 |
| N144_L528_P12_R6000_G180 | 144 | 528 | 12 | 6000 | 180 | $26892\text{s} \approx 7.47\text{h}$ | 0 |
| N144_L528_P12_R9000_G180 | 144 | 528 | 12 | 9000 | 180 | $36615\text{s} \approx 10.17\text{h}$ | 0 |
| N144_L528_P12_R12000_G180 | 144 | 528 | 12 | 12000 | 180 | $> 24\text{h}$ | - |

and utilising different optimisation solvers.

The computational burden of the proposed model arises from the rapid increase in the number of variables and constraints within the time-space network framework. Particularly, the significant rise in the number of integer variables, such as $PF^{r}_{i_{t_1} j_{t_2}}$ and $X_{i_{t_1} j_{t_2}}$, poses challenges for exact methods like branch-and-bound. To further reduce the computational complexity, the following measures can be adopted: (1) employing a rolling-horizon framework to divide the optimisation period into smaller horizons and subsequently resolving the model within each of these horizons; (2) clustering requests based on their spatial and temporal information; however, this approach might compromise accuracy for optimality; (3) developing tailored algorithms to tackle the issue, such as decomposition-based algorithms or meta-heuristics. It is worth mentioning that all these measures come with the potential drawback of losing optimal solutions.

## 4.7 Conclusions and future research

In this chapter, we propose a non-convex non-linear mathematical programming model to optimise fleet sizing and management decisions of an SAV service while considering traffic congestion and the non-linear demand of multi-class users (according to income). The congestion effect is measured through a dynamically varying travel time with respect to the traffic flow. Travellers' mode choice behaviour is modelled between SAVs and bicycles, assuming that no private cars are allowed in cities, which is

captured through an endogenous binary logit model. The two accept/rejection mechanisms (mandatory vs. non-mandatory acceptance) are explored, and the service level is endogenously determined which can affect travellers' willingness to use SAV services. The computational challenge posed by the non-linear and non-convex nature of the model is addressed through reformulation and the use of outer-inner approximation methods combined with a breakpoint generation algorithm to obtain a relaxed version of the original problem. The reformulated model can be solved using state-of-art solvers, such as Gurobi.

A quasi-real case study of Delft, in The Netherlands, was performed and a sensitivity analysis was carried out to demonstrate the performance of the proposed model and provide managerial insights to SAV operators in a promising future scenario. Results indicated that demand for SAVs, supply strategies of SAV operators, and network performance (traffic congestion) are interdependent with each other. Thus, it is crucial to take their interactions into account when managing fleets in an SAV service system. In terms of the fleet sizing strategy, computational results indicated that the initial distribution of the SAV operator's fleet is greatly impacted by factors such as the population's geographical distribution, land use patterns, and residents' travel behaviour. In addition, the fleet sizing decision is significantly influenced by the pricing strategy, unit operating costs of the SAV fleet, network congestion level, and the value of the parameters $\beta_0$. When the price rate is low, the fleet sizing decision is also sensitive to the accept/reject mechanism (mandatory vs. non-mandatory acceptance) and the travellers' sensitivity to the service quality level described by parameter $\beta_1$. The fleet sizing decision is insensitive to the change in the delay penalty. When the pricing rate of using SAVs is high, the fleet sizing is insensitive to parameter $\beta_1$. In addition, a low price of SAV service will attract more users but it may not necessarily bring a higher profit because of the increased traffic congestion. Besides, bringing fleets with lower operational costs to the system may earn more profits for an SAV operator through operational cost savings, reduction in delay penalties due to the improved traffic congestion, and lower depreciation costs of their fleets as less fleet is needed, despite the fact that SAVs had to take more detours and relocations.

Results indicate that SAV services are more attractive to travellers with a higher VOTT than those with a lower VOTT. Besides, travellers with a high VOTT are more sensitive to variations in trip length compared with the other classes. For long trips, travellers with high VOTT always prefer SAV services. However, for those with lower VOTT, SAV services are only preferred when the price is low. For middle and short trips, bicycles are preferable in most cases unless the price rate is low.

As a direction for future research, we propose the integration of the following aspects into our model: (1) demand and departure time stochasticity; (2) optimising the

pricing strategies; (3) worst-case scenarios in robust optimisation; (4) incorporation of ride-sharing mechanisms within the SAV service system; (5) interaction between SAVs and public transit systems.

# Appendix

# 4.A Problem formulation

We summarise the complete problem formulation in Chapter 4 as well as the notations of the sets, parameters, and variables below.

*Table 4.A.1: Notation of the sets, parameters, and variables*

| Notation | Description |
|---|---|
| **Set** | |
| $T$ | $= \{0, 1, 2, \ldots, \mathscr{T}\}$. Set of time instants in the operation period. |
| $N$ | Set of nodes. |
| $L$ | Set of road links between nodes in set $N$. |
| $G$ | Set of links in the time-space network. |
| $N_P$ | Set of nodes allowing parking for SAVs with $N_P \subseteq N$. |
| $R$ | Set of groups of trips, where each group of trips $r \in R$ has the same origin, destination, departure time, and latest arrival time at the destination. |
| $M$ | Set of travel modes, with the automated vehicles $(AV)$ and bicycles $(B)$ as the two options. |
| $K$ | $= \{1, 2, \ldots, k, \ldots, \mathscr{K}\}$. Index set of predetermined breakpoints. |
| **Parameters** | |
| $\Delta t$ | Time step. |
| $l_{ij}$ | Length of road link $(i, j) \in L$. |
| $Q_{ij}$ | Capacity of road link $(i, j) \in L$ in vehicles per time step. |
| $t_{ij}^{\max}$ | Maximum travel time by cars on road link $(i, j) \in L$. |
| $t_{ij}^{\min}$ | Minimum travel time by cars on road link $(i, j) \in L$. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of road link $(i, j) \in L$ in vehicles that fit on the road link from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |
| $\alpha$ | Trip service rate when all the requests have to be accepted, %. |
| $o^r$ | Origin node for group of trips $r \in R$. |
| $d^r$ | Destination node for group of trips $r \in R$. |
| $a^r$ | Departure time for group of trips $r \in R$. |
| $b^r$ | Latest arrival time for group of trips $r \in R$. |
| $sd^r$ | Shortest travel distance for group of trips $r \in R$, in kilometres. |
| $st^r$ | Shortest travel time assuming free-flow speed for group of trips $r \in R$, in time steps. |
| $n^r$ | Total number of trips for group $r \in R$. |
| $V_B^r$ | Deterministic systematic component of the utility of bicycles for group of trips $r \in R$. |
| $OM_m^r$ | Monetary costs of travellers in group $r \in R$ using mode $m \in M$, in euros. |
| $\beta_0$ | Parameter converting costs into utility, utility/euro. |
| $\beta_1$ | Parameter converting service rate into utility. |
| $\beta_m^r$ | Travellers' value of travel time in group $r$ using mode $m \in M$, euros/time step. |
| $T_B^r$ | Travel time of using bicycles for trips in group $r \in R$. |
| $p^0$ | Initial base fare for using SAVs, euros/trip. |
| $p$ | Travel distance-related price for using an SAV, euros/km. |
| $co$ | Unit driving operational cost of an SAV, euros/km. |
| $cd$ | Penalty for drop-off delay of passengers, euros/time step. |

| | |
|---|---|
| $cf$ | Depreciation cost in one hour for using an SAV, euros/vehicle . |
| $(u^k, \ln u^k)$ | Coordinates of the $k^{\text{th}}$ breakpoint. |
| $\mathscr{M}_r^1$ | Big-M parameter, where $r \in R$. |
| $\mathscr{M}_{ijt_1t_2}^2$ | Big-M parameter, where $t_1, t_2 \in T$, if $t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min}, (i,j) \in L$. |

**Decision variables**

| | |
|---|---|
| $V_{AV}^r$ | Deterministic systematic component of travellers' utility for using an SAV in group $r \in R$. |
| $T_{AV}^r$ | Longest SAVs travel time for group $r \in R$. |
| $P_{AV}^r$ | Probability to choose SAVs for the trips in group $r \in R$. |
| $D_{AV}^r$ | Total number of trips using SAVs in group $r \in R$. |
| $\alpha$ | Trip service rate when some requests can be rejected. |
| $V$ | SAV fleet size. |
| $V_i$ | Initial distribution of SAVs at parking node $i \in N_p$ at the beginning of a day. |
| $S^r$ | Total number of trips served by SAVs from group $r$, where $r \in R$. |
| $PF_{i_{t_1} j_{t_2}}^r$ | Passenger flow in the group of trips $r \in R$ served by an SAV in road link $(i,j)$, from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G, a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| $E_t^r$ | Total number of passengers in group of trips $r \in R$ arriving at time $t \in T$. |
| $F_{i_{t_1} j_{t_2}}$ | Vehicle flow in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. Note that when $t_1 = 0$, $i \in N_P$, meaning that SAVs have to depart from the parking nodes at the beginning of a day. |
| $W_{i_t}$ | Total number of SAVs parking at node $i \in N_P$ from time instant $t$ to $t+1$, with $t \in T$. |
| $Z_t^r$ | Binary variable with $r \in R, t \in T$ if $a^r + st^r \leq b^r$. |
| $X_{i_{t_1} j_{t_2}}$ | Binary variable which is 1 when any vehicle travels in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. |
| $A_t^r$ | Binary variable which is 1 when at least one trip in group $r \in R$ arrives at time $t \in T$, and 0 otherwise. |
| $LN_{AV}^r$ | Auxiliary continuous variable, where $r \in R$. |
| $LN_B^r$ | Auxiliary continuous variable, where $r \in R$. |
| $\lambda_r^k$ | Binary variable indicating whether an interval $[u^k, u^{k+1}]$ is active or not, where $k \in \{1,2,\ldots,k,\ldots,\mathscr{K}-1\}$, $r \in R$. |
| $\theta_r^k$ | Convex combination coefficient for breakpoint $k \in K$ for group of trips $r \in R$. |
| $\overline{\lambda}_r^k$ | Binary variable indicating whether an interval $[1-u^{k+1}, 1-u^k]$ is active or not, where $k \in \{1,2,\ldots,k,\ldots,\mathscr{K}-1\}$, $r \in R$. |
| $\overline{\theta}_r^k$ | Convex combination coefficient for breakpoint $k \in K$ for group of trips $r \in R$. |
| $\overline{D}_h$ | Binary variables utilised for discretising integer variables, where $h \in \{0,1,\ldots,\mathscr{H}\}$. |
| $Y_h$ | Continuous variables utilised for describing the value of the integer variables, where $h \in \{0,1,\ldots,\mathscr{H}\}$. |

*Mixed integer linear program*

$$\max \sum_{r\in R} OM_{AV}^{r} S^{r} - cf \cdot V - co\left(\sum_{(i_{t_1}, j_{t_2})\in G} l_{ij} F_{i_{t_1} j_{t_2}}\right) - cd \sum_{r\in R}\left(\sum_{t\in T} tE_{t}^{r} - a^{r} S^{r} - st^{r} S^{r}\right) \tag{4.67}$$

where

$$OM_{AV}^{r} = p^{0} + sd^{r} p, \qquad \forall r \in R. \tag{4.68}$$

subject to:

$$V_{AV}^{r} = -\beta_0 (OM_{AV}^{r} + \beta_{AV}^{r} T_{AV}^{r}) - \beta_1 (1-\alpha), \qquad \forall r \in R \tag{4.69}$$

$$n^{r} P_{AV}^{r} - 0.5 < D_{AV}^{r} \leq n^{r} P_{AV}^{r} + 0.5, \qquad \forall r \in R \tag{4.70}$$

$$\sum_{r\in R} D_{AV}^{r} = \sum_{h=0}^{\mathscr{H}} 2^{h} \overline{D}_{h} \tag{4.71}$$

$$Y_h \leq \alpha, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.72}$$

$$Y_h \leq \overline{D}_h, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.73}$$

$$Y_h \geq \alpha + \overline{D}_h - 1, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.74}$$

$$\sum_{h=0}^{\mathscr{H}} 2^{h} Y_h = \sum_{r\in R} S^{r} \tag{4.75}$$

$$S^{r} \leq D_{AV}^{r}, \qquad \forall r \in R \tag{4.76}$$

$$S^{r} = \sum_{j_t | (o_{a^r}^{r}, j_t)\in G} PF_{o_{a^r}^{r} j_t}^{r}, \qquad \forall r \in R \tag{4.77}$$

$$S^{r} = \sum_{t\in T | a^{r}+st^{r} \leq t \leq b^{r}} E_{t}^{r}, \qquad \forall r \in R \tag{4.78}$$

$$E_{t}^{r} = \sum_{i_{t_1} | (i_{t_1}, d_t^{r})\in G} PF_{i_{t_1} d_t^{r}}^{r}, \qquad \forall r \in R, t \in T \tag{4.79}$$

$$\sum_{j_{t_1} | (d_t^{r}, j_{t_1})\in G} PF_{d_t^{r} j_{t_1}}^{r} = 0, \qquad \forall r \in R, a^{r} \leq t \leq b^{r} \tag{4.80}$$

$$\sum_{i_{t_1} | (i_{t_1}, o_t^{r})\in G} PF_{i_{t_1} o_t^{r}}^{r} = 0, \qquad \forall r \in R, a^{r} \leq t \leq b^{r} \tag{4.81}$$

$$\sum_{j_{t_0}|(j_{t_0},i_{t_1})\in G} PF^r_{j_{t_0}i_{t_1}} = \sum_{j_{t_2}|(i_{t_1},j_{t_2})\in G} PF^r_{i_{t_1}j_{t_2}}, \qquad \forall r\in R, a^r < t_1 < b^r, i\in N, i\neq o^r, i\neq d^r \tag{4.82}$$

$$\sum_{r\in R} PF^r_{i_{t_1}j_{t_2}} \leq F_{i_{t_1}j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2})\in G \tag{4.83}$$

$$\sum_{j_{t_1}|(j_{t_1},i_t)\in G, t_1<t} F_{j_{t_1}i_t} = \sum_{j_{t_2}|(i_t,j_{t_2})\in G, t<t_2} F_{i_t j_{t_2}}, \qquad \forall i\in N\setminus N_P, 0<t<\mathscr{T}, \tag{4.84}$$

$$\sum_{j_{t_1}|(j_{t_1},i_t)\in G, t_1<t} F_{j_{t_1}i_t} + W_{i_{t-1}} = \sum_{j_{t_2}|(i_t,j_{t_2})\in G, t<t_2} F_{i_t j_{t_2}} + W_{i_t}, \qquad \forall i\in N_P, 0<t<\mathscr{T}, \tag{4.85}$$

$$\sum_{j_t|(i_0,j_t)\in G} F_{i_0 j_t} + W_{i_0} = V_i, \qquad \forall i\in N_p \tag{4.86}$$

$$\sum_{i\in N_p} V_i = V \tag{4.87}$$

$$\frac{E^r_t}{n^r} \leq A^r_t \leq E^r_t, \qquad \forall r\in R, a^r + st^r \leq t \leq b^r \tag{4.88}$$

$$T^r_{AV} \geq A^r_t (t-a^r), \qquad \forall r\in R, a^r + st^r \leq t \leq b^r \tag{4.89}$$

$$T^r_{AV} \leq A^r_t (t-a^r) + (b^r - a^r)(1-Z^r_t), \qquad \forall r\in R, a^r + st^r \leq t \leq b^r \tag{4.90}$$

$$\sum_{t|a^r+st^r\leq t\leq b^r} Z^r_t = 1, \qquad \forall r\in R \tag{4.91}$$

$$\sum_{t_2|(i_{t_1},j_{t_2})\in G} X_{i_{t_1}j_{t_2}} \leq 1, \qquad \forall (i,j)\in L, t_1\in T \tag{4.92}$$

$$F_{i_{t_1}j_{t_2}} \leq \left\lfloor C_{i_{t_1}j_{t_2}} \right\rfloor X_{i_{t_1}j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2})\in G \tag{4.93}$$

$$t_1 + \sum_{t\in T} X_{i_{t_1}j_t}(t-t_1) \leq t_2 + \sum_{t\in T} X_{i_{t_2}j_t}(t-t_2) + (t_1 + t^{\max}_{ij} - t_2)\left(1 - \sum_{t\in T} X_{i_{t_2}j_t}\right), \quad \forall (i,j)\in L, t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij} \tag{4.94}$$

$$LN^r_{AV} - LN^r_B = V^r_{AV} - V^r_B, \qquad \forall r\in R \tag{4.95}$$

$$LN^r_{AV} \leq \frac{1}{u^k} P^r_{AV} + \ln u^k - 1, \qquad \forall r\in R, k\in K \tag{4.96}$$

$$LN^r_{AV} \geq \sum_{k=1}^{\mathscr{K}} \theta^k_r \ln u^k, \qquad \forall r\in R \tag{4.97}$$

$$P_{AV}^{r} = \sum_{k=1}^{\mathscr{K}} \theta_r^k u^k, \qquad \forall r \in R \tag{4.98}$$

$$\sum_{k=1}^{\mathscr{K}} \theta_r^k = 1, \qquad \forall r \in R \tag{4.99}$$

$$\sum_{k=1}^{\mathscr{K}-1} \lambda_r^k = 1, \qquad \forall r \in R \tag{4.100}$$

$$\theta_r^1 \leq \lambda_r^1, \qquad \forall r \in R \tag{4.101}$$

$$\theta_r^k \leq \lambda_r^{k-1} + \lambda_r^k, \qquad \forall r \in R, k \in \{2, \ldots, \mathscr{K}-1\} \tag{4.102}$$

$$\theta_r^{\mathscr{K}} \leq \lambda_r^{\mathscr{K}-1}, \qquad \forall r \in R \tag{4.103}$$

$$LN_B^r \leq \frac{1}{u^k}(1 - P_{AV}^r) + \ln u^k - 1, \qquad \forall r \in R, k \in K \tag{4.104}$$

$$LN_B^r \geq \sum_{k=1}^{\mathscr{K}} \overline{\theta}_r^k \ln u^k, \qquad \forall r \in R \tag{4.105}$$

$$1 - P_{AV}^r = \sum_{k=1}^{\mathscr{K}} \overline{\theta}_r^k u^k, \qquad \forall r \in R \tag{4.106}$$

$$\sum_{k=1}^{\mathscr{K}} \overline{\theta}_r^k = 1, \qquad \forall r \in R \tag{4.107}$$

$$\sum_{k=1}^{\mathscr{K}-1} \overline{\lambda}_r^k = 1, \qquad \forall r \in R \tag{4.108}$$

$$\overline{\theta}_r^1 \leq \overline{\lambda}_r^1, \qquad \forall r \in R \tag{4.109}$$

$$\overline{\theta}_r^k \leq \overline{\lambda}_r^{k-1} + \overline{\lambda}_r^k, \qquad \forall r \in R, k \in \{2, \ldots, \mathscr{K}-1\} \tag{4.110}$$

$$\overline{\theta}_r^{\mathscr{K}} \leq \overline{\lambda}_r^{\mathscr{K}-1}, \qquad \forall r \in R \tag{4.111}$$

$$0 \leq \alpha \leq 1 \tag{4.112}$$

$$V_{AV}^r \geq 0, \qquad \forall r \in R \tag{4.113}$$

$$T_{AV}^r \in \mathbb{N}^0, \qquad \forall r \in R \tag{4.114}$$

$$P_{AV}^r \geq 0, \qquad \forall r \in R \tag{4.115}$$

$$D_{AV}^r \in \mathbb{N}^0, \qquad \forall r \in R \tag{4.116}$$

$$\overline{D}_h \in \{0,1\}, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.117}$$

$$Y_h \geq 0, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{4.118}$$

$$V \in \mathbb{N}^0 \tag{4.119}$$

$$V_i \in \mathbb{N}^0, \qquad \forall i \in N_P \tag{4.120}$$

$$S^r \in \mathbb{N}^0, \qquad \forall r \in R \tag{4.121}$$

$$E_t^r \in \mathbb{N}^0, \qquad \forall r \in R, t \in T \tag{4.122}$$

$$PF_{i_{t_1} j_{t_2}}^r \in \mathbb{N}^0, \qquad \forall r \in R, (i_{t_1}, j_{t_2}) \in G \tag{4.123}$$

$$F_{i_{t_1} j_{t_2}} \geq 0, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{4.124}$$

$$W_{i_t} \geq 0, \qquad \forall i \in N_P, t \in T \tag{4.125}$$

$$Z_t^r \in \{0,1\}, \qquad \forall r \in R, t \in T, a^r + st^r \leq b^r \tag{4.126}$$

$$X_{i_{t_1} j_{t_2}} \in \{0,1\}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{4.127}$$

$$A_t^r \in \{0,1\}, \qquad \forall r \in R, t \in T, a^r + st^r \leq t \leq b^r \tag{4.128}$$

$$LN_{AV}^r \geq 0, \qquad \forall r \in R \tag{4.129}$$

$$LN_B^r \geq 0, \qquad \forall r \in R \tag{4.130}$$

$$\theta_r^k \geq 0, \qquad \forall r \in R, k \in K \tag{4.131}$$

$$\lambda_r^k \in \{0,1\}, \qquad \forall r \in R, k \in K \tag{4.132}$$

$$\overline{\theta}_r^k \geq 0, \qquad \forall r \in R, k \in K \tag{4.133}$$

$$\overline{\lambda}_r^k \in \{0,1\}, \qquad \forall r \in R, k \in K \tag{4.134}$$

# Chapter 5

# Solution methods for pricing and fleet management in shared automated vehicle services considering supply-demand dynamics, congestion, and income heterogeneity

---

Research on supply-demand dynamics in Shared Automated Vehicles (SAVs) services has grown rapidly. However, developed models are often complex non-linear systems that face significant challenges when solving. This chapter aims to: (1) investigate optimal pricing and fleet management strategies for SAV services, considering the interplay between demand and supply side variables, congestion effects, and the heterogeneous income levels of travellers; (2) propose various solution methods and conduct a comparative analysis of these methods. For modelling the problem, we propose a mixed-integer non-linear programming model under three different pricing strategies: base fare plus distance-based fare, distance-based fare only, and income class-based fare. For solving the problem, we present three distinct solution algorithms that tackle the model's complex non-linearities. These consist of using linearisation techniques, a hybrid method based on Particle Swarm Optimisation (PSO), and a hybrid method based on Bayesian Optimisation (BO).

This chapter is structured as follows: Section 5.1 provides an introduction to the background. Section 5.2 reviews the existing literature on pricing strategies in ride-hailing services and the literature that models the dynamics of supply-demand inter-

actions. Section 5.3 describes the non-linear, non-convex mathematical model formulated for this study. Section 5.4 presents the three proposed solution algorithms, explaining each method comprehensively. In Section 5.5, to illustrate the practical application of the models, we first conduct a case study on a small-scale problem, followed by a quasi-real case study of the city of Delft in the Netherlands. Finally, Section 5.6 concludes the chapter with a summary of the main findings and suggestions for further research.

## 5.1 Introduction

Supply-demand interactions in the ride-hailing market are an important element that should not be ignored when making policy, management and operational decisions, especially when it comes to researching future mobility systems—such as Shared Automated Vehicle (SAV) services—where the demand is unknown. Such a future mobility system consists of a variety of components that depend on each other interactively. From the perspective of the SAV service suppliers, decisions need to be made at the planning and operational levels. Planning decisions, made before launching the SAV service, might include aspects like pricing strategy, SAV fleet sizing, initial fleet distribution, and service quality. Operational decisions, made in response to new SAV service requests and real-time network conditions, may involve trip assignment, vehicle routing, relocation, parking, and decisions to accept or reject requests. These decisions, in turn, influence passengers' choice to use the service.

On the demand side of the SAV service system, passenger choices are notably influenced by factors such as price, travel time, and service quality. Moreover, heterogeneous passengers with diverse socio-demographic characteristics—including variations in income level, age, and gender—demonstrate different sensitivities to these factors. In such a SAV service system, the decisions on the supply side are inherently linked to those on the demand side. However, many studies tend to assume that demand is predetermined when optimising supply-side decisions, overlook the time-varying network conditions or fail to consider the heterogeneous preferences of passengers. This triggers the need for modelling the complex interactions of these decision variables dynamically. Developing such models is crucial for enabling service suppliers to make profitable and realistic decisions in the management and operations of current and future systems.

Among all the planning decisions, pricing exerts the most direct and significant influence on passengers' willingness to use the service (Chen et al., 2020). Pricing is an effective tool not only for balancing supply and demand but also for alleviating traffic congestion. Consequently, understanding demand patterns and designing effective pricing strategies has attracted considerable interest from both academia and the industry. Although various pricing strategies and fare structures have been explored, few studies have investigated optimal pricing strategies tailored to diverse demographic groups. In this chapter, we investigate three pricing strategies: (1) a base fare plus a distance-based fare, (2) a solely distance-based fare, and (3) an income class-based fare. Income class-based fares devise prices according to passengers' income levels, making SAV services more accessible to low-income individuals and potentially boost-

ing overall usage and social equity (Verbich & El-Geneidy, 2017; Dong et al., 2022). Moreover, charging higher fares to high-income users could balance the service's overall affordability and enhance sustainability.

In this chapter, we explore the optimal pricing and fleet management challenges for an SAV service provider that operates a fleet providing ride-hailing services in a prospective mobility system. In this envisioned system, travellers have two transportation options: SAVs and bicycles. This scenario is anticipated as cities increasingly advocate for the elimination of private cars to create car-free environments, as discussed in studies of Nieuwenhuijsen & Khreis (2016) and Fan et al. (2023). Travellers use smartphone apps to request SAVs by entering trip details. The service platform evaluates these requests, accepting them if beneficial to the company or rejecting them if otherwise, in which case travellers opt for bicycles if they offer greater utility. Accepted requests are matched with available SAVs, which are dispatched to pick up customers.

The mathematical model proposed in this chapter is a Mixed-integer Nonlinear Programming (MINLP) model adapted from the model by Fan et al. (2023), with two major modifications. Firstly, three pricing strategies are modelled additionally, with pricing decisions as endogenous variables. Secondly, the diverse sensitivities of different passenger income classes are incorporated into the demand modelling to explore their impact on these pricing strategies. In addition, the endogenous demand for SAVs is modelled using a binary logit model, where price, travel time, and service quality influence travellers' willingness to use SAV services. Traffic congestion is modelled by dynamically varying travel time based on vehicle flow on road links. The accept/reject decisions, which reflect service quality, are also endogenously determined in the model.

With the modifications to the model, its complexity has increased due to the non-linearity introduced by both non-linear constraints and a non-linear objective function, as well as the addition of new variables. These changes pose significant challenges for conventional solution methods. To address this complexity, the existing literature typically employs but is not limited to, the following strategies: (1) linearisation techniques (Guo et al., 2022; Fan et al., 2023), which convert non-linear components into a form that can be solved using linear programming, thereby simplifying the complexity of the original problem; (2) reformulation techniques (Huang et al., 2018; Lu et al., 2021), which reformulate the problem, identify problem structure or decompose the problem into smaller and more manageable problems; (3) heuristic/metaheuristic algorithms that seek to find sufficiently good solutions for complex optimisation challenges where exact methods are impractical (Huang et al., 2018; Lu et al., 2021).

In response to these challenges, our research has developed three tailored solu-

tion methods. The first method utilises linearisation techniques. The second employs a hybrid approach that combines Particle Swarm Optimisation (PSO), a metaheuristic, with reformulation techniques. The third method employs Bayesian Optimisation (BO) alongside linearisation techniques. BO is particularly effective in contexts where evaluating the objective function is costly or complex, such as in parameter tuning for machine learning algorithms. In our study, it plays a key role in simplifying the objective function of the mathematical model. To enhance computational efficiency, the second and third methods are implemented in parallel. A comparative analysis of these methodologies is conducted to assess their effectiveness and efficiency in solving the proposed model.

The main contribution of this chapter is therefore twofold: (1) We propose an adapted model which investigates three pricing strategies and fleet management decisions, incorporating supply-demand interactions, congestion and the heterogeneous demographic characteristics of passengers; (2) We develop and compare three different solution algorithms in a comparative study aimed at identifying the method that best balances performance and computational efficiency.

## 5.2 Literature review

The literature review is divided into two main sections: Section 5.2.1 focuses on the existing literature addressing pricing problems in ride-hailing services; and Section 5.2.2 dives into the literature which models the endogenous supply-demand interactions, presenting the demand modelling techniques and main methodologies. This review aims to provide a comprehensive understanding of both pricing problems and supply-demand modelling in the context of ride-hailing services.

### 5.2.1 Pricing problems in ride-hailing services

Pricing problems in ride-hailing services have gained a wide range of research interests (Tong et al., 2018; Wang et al., 2016; Nourinejad & Ramezani, 2020; Zhang & Nie, 2021; Asadpour et al., 2023). Existing research on the practice of ride-hailing services focuses on a two-sided market, where both the fare paid by passengers and the wage paid to drivers need to be determined. However, in the context of using shared automated taxis for ride-hailing services, the problem is simplified because there is no need to hire human drivers anymore. Thus, we only focus on the pricing strategies for the fare paid by the passengers.

Pricing mechanisms can be categorised into two types: static pricing and dynamic

pricing. Static pricing strategies determine a fixed price structure throughout the planning period. Research has investigated, amongst others, distance-based pricing (Dong et al., 2022), travel time-based pricing (Liang et al., 2020), origin- and/or destination-based pricing (Özkan, 2020; Müller et al., 2023), zone-based spatial pricing (Li et al., 2021; Dong et al., 2022). Dynamic pricing allows the price and/or wage to vary according to various factors such as the current market demand and supply relationship, time, or even competitor pricing. Examples of this include surge pricing (Al-Kanj et al., 2020), spatial-temporal pricing (Meskar et al., 2023), and congestion pricing (Zheng et al., 2023). Surge pricing mechanisms usually charge passengers a higher price during peak hours, extreme weather conditions, or major events where the supply-demand is imbalanced.

Existing research on static and dynamic pricing presents a range of controversial views. Nourinejad & Ramezani (2020) address three pricing strategies, which are static pricing, constrained dynamic pricing (assuming the instantaneous profit is also non-zero), and unconstrained dynamic pricing (allowing the instantaneous profit to be non-zero). Their results show that the unconstrained dynamic pricing strategy provides the highest overall profit and ensures a more stable rider waiting time with less variation during the study period. Similarly, Cachon et al. (2017) find that incorporating dynamic elements into pricing and/or wages can enhance profits compared to fixed pricing strategies. However, despite its profitability, surge pricing has faced criticism for its potential negative impact on consumer welfare, as it forces consumers to either pay higher prices or wait longer during peak demand periods (Ashkrof et al., 2022a). Moreover, Lin & Zhou (2019) suggest that surge pricing is not always the optimal choice for ride-hailing companies, as static pricing can achieve comparable performance. In this study, we focus on investigating static pricing strategies during surge demand periods (peak hours) in a typical working day. During this optimisation period, the pricing strategy stays the same.

People with different socio-demographic characteristics—such as gender, age and income levels—display different sensitivities to price changes. The behaviour of heterogeneous users has a great impact on the platform's optimal pricing and wage decisions (Taylor, 2018; Wu et al., 2020). Beirigo et al. (2022) segment users of an autonomous mobility-on-demand system into two classes: first-class users, who are willing to pay more for a higher level of service, and second-class users. Bai & Tang (2022) proposed a model to study the equilibrium between two competing on-demand service platforms. Their proposed model can capture some market characteristics such as the price- and time-sensitive customers and earning-sensitive service providers. Dong et al. (2022) introduced class-based pricing strategies in a chance-constrained dial-a-ride problem where heterogeneous users are grouped into classes

by their socio-demographic characteristics. However, the pricing is introduced as a class-based parameter which is not endogenously determined. In the existing pricing problems, very few have considered an income class-based fare considering people's heterogeneous sensitivity toward the costs/money. Thus, in our work, we plan to study the optimal income class-based fare and its interplay between this pricing variable and all the other planning and operational level variables.

### 5.2.2 Endogenous supply-demand interaction modelling

The supply-demand interaction has been studied in transportation systems for different problem settings, for example, in the two-sided market of ride-hailing services (Li et al., 2021; Meskar et al., 2023), carsharing systems (Huang et al., 2018; Lu et al., 2021), dial-a-ride problems (Dong et al., 2022), fleet sizing and management problems using SAVs (Guo et al., 2022; Fan et al., 2023), public transit planning problems (Cadarso et al., 2017; Steiner & Irnich, 2020; Wei et al., 2022), and competition among multiple on-demand mobility services (Wang et al., 2022b). In Table 5.2.1, we list several papers that consider supply-demand interactions with their main decisions, methodologies, and how demand is modelled.

Accurately modelling the interactions between decision variables on both the supply and demand sides necessitates that demand is determined endogenously. Normally, demand can be modelled by (a) a simple linear or non-linear function such as an exponential function (Jorge et al., 2015; Huang et al., 2020); (b) a discrete choice model, such as a binary logit model (Lu et al., 2021; Guo et al., 2022; Fan et al., 2023), a multinomial logit model (Zhang & Nie, 2021), a nested logit model (Cadarso et al., 2017), a dynamic discrete choice model (Västberg et al., 2020), or chance-based constraints (Dong et al., 2022); (c) a simulation-based approximation of the discrete choice model (Paneque et al., 2021, 2022). The demand model is then integrated into optimisation models (Huang et al., 2018; Lu et al., 2021; Paneque et al., 2021; Zhang & Nie, 2021; Paneque et al., 2022; Fan et al., 2023), simulation-based methods (Hörl et al., 2021; Wang et al., 2022b), or a hybrid of both (Pinto et al., 2020).

Incorporating demand modelling into an optimisation problem often presents challenges to the solution process, typically requiring the development of tailored solution algorithms. Some papers use aggregated market equilibrium models (Nourinejad & Ramezani, 2020; Li et al., 2021; Zhang & Nie, 2021). However, these models fail to capture temporal dynamics (such as departure time and arrival time). In our problem, it is crucial to account for detailed operational decisions, such as trip assignment and vehicle routing, to accurately reflect the dynamically varying congestion effects. Simulation-based methods excel at capturing microscopic details by reconstructing

complex scenarios. However, they often suffer from long computational times, particularly when there are many decision variables to be combined. Therefore, they are often used to assess multiple scenarios instead of searching for an optimal combination. Different from most existing studies, we develop optimisation-based methods that not only describe the supply-demand endogenous interactions but also capture the detailed operational decisions of SAVs and the dynamic effects of network congestion. Due to the complexity of the problem, three tailored solution algorithms are designed to address the model efficiently.

## 5.3 Problem formulation

In this section, we first present the main assumptions in Section 5.3.1. Then, we describe the mathematical model, including model settings in Section 5.3.2, passenger demand modelling in Section 5.3.3, SAV service planning and operation in Section 5.3.4, traffic congestion modelling in Section 5.3.5, and the objective function in Section 5.3.6.

### 5.3.1 Assumptions

Some assumptions are made in this chapter: (a) It is assumed that the total mobility demand within an urban area is fixed and predetermined. (b) The SAVs in this study operate at SAE level 5 (On-Road Automated Driving (ORAD) committee, 2021), allowing them to navigate the entire network autonomously without a human driver. (c) The use of privately owned or human-driven vehicles is not included in this study. (d) Travellers are assumed to use only one mode of transportation; mode transfers are not considered. (e) Pooled service options are excluded from consideration in this research.

### 5.3.2 Model setting

The problem is formulated as a mathematical model aiming at optimising planning decisions by modelling the interactions between passenger demand, TNC supply, and traffic congestion. Detailed decisions and their relationships are depicted in Figure 5.3.1.

We consider peak-hour traffic on a typical workday in an urban area by aggregating trips with identical travel information and socio-demographic information into groups. The travel data includes origin, destination, departure time, and latest arrival time.

*Table 5.2.1: Selected papers on supply-demand interaction in transportation systems.*

| Transport mode or service | Authors | Main decisions | Methodology | Demand modelling |
|---|---|---|---|---|
| Carsharing system competing with private cars | Huang et al. (2018) | Carsharing station location, station capacity, fleet size, vehicle relocation, and mode choice | A mixed-integer nonlinear programming model, solved by a customised gradient algorithm | Modelled using a binary logit model |
| Two-sided ride-sourcing market | Nourinejad & Ramezani (2020) | Passenger demand, drivers supply, fare for riders, wage for drivers, passenger waiting time, driver cruising time, | A macroscopic analytical model and a microscopic simulation model used within a model predictive control approach | Modelled as a function of generalised cost |
| Transit networks and shared autonomous mobility fleets | Pinto et al. (2020) | Transit frequency, fleet size of SAVs, mode choice, SAV route choice, traffic congestion | A bi-level mathematical programming formulation, solved by a heuristic procedure combining optimisation and simulation approaches | Modelled using a multinomial logit model |
| Carsharing system competing with private cars | Lu et al. (2021) | Price, vehicle relocation, mode choice | A bilevel nonlinear mathematical programming model, solved by genetic algorithm | Modelled using a binary logit model |
| Automated taxis | Hörl et al. (2021) | Price, waiting time, fleet size, demand | An agent-based simulation | Modelled using a multinomial logit model |
| E-hailing services integrated with transit service | Zhang & Nie (2021) | Mode choice, passenger demand, vehicle supply, pricing strategies for riders, wage for drivers, passenger waiting time | An aggregate equilibrium model, solved using a simple iterative fixed-point algorithm | Modelled using a multinomial logit model |
| Dial-a-ride problem | Dong et al. (2022) | Fleet size, accept/reject decisions, trip assignment, and vehicle routing | A mixed-integer programming model, solved by two customised heuristic algorithms | Modelled using logit-based chance constraints |
| Public transit planning integrated with ride-hailing | Wei et al. (2022) | Public transit schedule and operations, mode choices, ride-hailing operations, traffic congestion | A mixed-integer nonlinear programming model, solved by a bilevel decomposition algorithm | Modelled using a multinomial logit model |
| SAV management | Guo et al. (2022) | Vehicle operation, mode choice | A mixed-integer nonlinear programming model, reformulated and then solved by a solver | Modelled using a binary logit model |
| Competition between multiple AMoD systems | Wang et al. (2022b) | Fleet sizes, assignment strategies, pricing strategies, mode choice | An agent-based simulation | Modelled using a multinomial logit model |
| SAV fleet sizing and management | Fan et al. (2023) | Fleet size, initial fleet distribution, mode choice, service quality, congestion, trip assignment, and vehicle operation | A mixed-integer nonlinear programming model, reformulated and then solved by a solver | Modelled using a binary logit model |

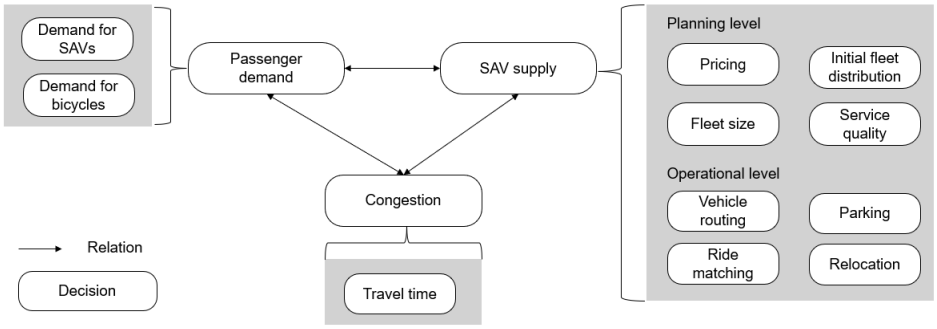



*Figure 5.3.1: Main decisions and their interactions.*

For socio-demographic information, we categorise travellers based on income levels—high, middle, and low. Travellers within each group have the freedom to choose their mode of transportation (either SAV or bicycle). Due to the congestion effects, the SAVs serving the travellers in the same groups are allowed to choose different routes to avoid the congestion. The sets and parameters used to detail this trip information are the following.

Sets associated with trips

| | |
|---|---|
| $IC$ | Set of income classes, categorised into low income $(low)$, middle income $(mid)$, and high income $(high)$. |
| $R^c$ | Set of groups of trips associated with a specific income class $c \in IC$. Each group $r \in R^c$ consists of trips that share the same characteristics, including origin, destination, departure time, latest arrival time, and income level $c$. |
| $M$ | Set of travel modes, consisting of automated vehicles $(AV)$ and bicycles $(B)$ as options. |

Parameters associated with trips

| | |
|---|---|
| $o^r$ | Origin node for group of trips $r \in R^c, c \in IC$. |
| $d^r$ | Destination node for group of trips $r \in R^c, c \in IC$. |
| $a^r$ | Departure time for group of trips $r \in R^c, c \in IC$. |
| $b^r$ | Latest arrival time for group of trips $r \in R^c, c \in IC$. |
| $sd^r$ | Shortest travel distance for group of trips $r \in R^c, c \in IC$, in kilometres. |
| $st^r$ | Shortest travel time assuming free-flow speed for group of trips $r \in R^c, c \in IC$, in time steps. |
| $n^r$ | Total number of trips for group $r \in R^c, c \in IC$. |

We employ a time-space network to address dynamic operational decision-making and the endogenous traffic congestion resulting from the large-scale deployment of SAVs. This network, an extension of the conventional directed network, spans multiple discrete time periods, represented as $T$, each with a duration of $\Delta t$. At every discrete time $t \in T$, the network is duplicated. The sets and parameters associated with the

network are shown as follows:

| Sets associated with the network setting | |
|---|---|
| $T$ | Set of discrete time instants $T = \{0, 1, 2, \ldots, \mathscr{T}\}$ in the operational period. |
| $N$ | Set of physical nodes within the network. |
| $L$ | Set of road links connecting the nodes in set $N$. |
| $G$ | Set of links in the time-space network. |
| $N_P$ | Subset of nodes that allow parking for SAVs, where $N_P \subseteq N$. |
| **Parameters associated with the network setting** | |
| $\Delta t$ | Time step. |
| $l_{ij}$ | Length of road link $(i, j) \in L$. |
| $Q_{ij}$ | Capacity of road link $(i, j) \in L$ in number of vehicles per time step. |
| $t_{ij}^{\max}$ | Maximum travel time by car on road link $(i, j) \in L$. |
| $t_{ij}^{\min}$ | Minimum travel time by car on road link $(i, j) \in L$. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of road link $(i, j) \in L$ in number of vehicles that fit on the road link from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |

### 5.3.3 Passenger demand modelling

We employ a discrete choice model to express travel mode preferences among travellers. Utility values for the two modes considered in this chapter—SAVs and bicycles—are computed. We use a binary logit model to estimate the probability of using each mode for different groups of trips. The parameters and variables relevant to this section are shown in Table 5.3.1.

**Binary logit model:**

The utility of using mode $m \in M$ for travellers in group $r \in R^c, c \in IC$ is expressed as $U_m^r = V_m^r + \varepsilon_m^r$. Here, $V_m^r$ represents the deterministic component of the utility, which is a function of observed attributes of the alternatives and characteristics of the individual. $\varepsilon_m^r$ is a random component reflecting all the unobservable influence on the utility (Ben-Akiva et al., 1985). Assuming $\varepsilon_m^r$ is independently and identically Gumbel distributed, we can obtain the probability $P_{AV}^r$ of using SAVs for groups $r \in R^c, c \in IC$ using a binary logit model, as shown in Equations (5.1). This equation is non-linear due to the presence of exponential terms in both the numerator and the denominator. To determine the probability of using SAVs, we need to know the deterministic utility $V_{AV}^r$ of SAVs and $V_B^r$ of bicycles, which can be expressed as Equation (5.2) and (5.3), respectively.

$$P_{AV}^r = \frac{e^{V_{AV}^r}}{e^{V_{AV}^r} + e^{V_B^r}}, \qquad \forall r \in R^c, c \in IC \tag{5.1}$$

The deterministic utility $V_{AV}^r$ of using SAVs includes three components as given in Equation 5.2: the fare for the services, travel time-related costs, and traveller satis-

*Table 5.3.1: Notation used for passenger demand modelling.*

| Notation | Description |
|---|---|
| Parameters | |
| $V_B^r$ | Deterministic systematic component of the utility of bicycles for group of trips $r \in R^c, c \in IC$. |
| $OM_B^r$ | Monetary costs of travellers in group $r \in R^c, c \in IC$ using bicycles, in euros. |
| $\beta_0^c$ | Parameter converting generalised costs into utility for income class $c \in IC$, in utility/euro. |
| $\beta_1$ | Parameter converting service rate into utility. |
| $VOT_m^c$ | Travellers' value of travel time in class $c \in IC$ using mode $m \in M$, in euros/time step. |
| $T_B^r$ | Travel time of using bicycles for trips in group $r \in R^c, c \in IC$. |
| Variables | |
| $P_{AV}^r$ | Probability to choose SAVs for trips in group $r \in R^c, c \in IC$. |
| $V_{AV}^r$ | Deterministic systematic component of travellers' utility for using an SAV in group $r \in R^c, c \in IC$. |
| $OM_{AV}^r$ | Monetary costs of travellers in group $r \in R^c, c \in IC$ using SAVs, in euros. |
| $T_{AV}^r$ | Longest SAVs travel time for group $r \in R^c, c \in IC$. |
| $\alpha$ | Trip service rate. |
| $D_{AV}^r$ | Total number of trips using SAVs in group $r \in R^c, c \in IC$. |

faction concerning the trip rejection rate. To capture the sensitivity of travellers from different income classes to changes in costs (measured in euros), we introduce a class-specific parameter, $\beta_0^c$. Individuals with higher incomes tend to perceive costs less negatively than those with middle or lower incomes, with middle-income individuals perceiving costs less negatively than those with low incomes. The cost for group $r \in R^c, c \in IC$ is determined as the monetary cost of using SAV services $OM_{AV}^r$ and the travel time-related cost $VOT_{AV}^c T_{AV}^r$. Here, $VOT_{AV}^c$ represents the income class-based value of travel time for using AVs, and $T_{AV}^r$ represents the travel time of using AVs by group $r \in R^c, c \in IC$. Trips can be rejected at a rate of $1-\alpha$. The sensitivity of travellers to this is described by the parameter $\beta_1$. Rejecting trips negatively impacts people's willingness to use SAV services. The deterministic term $V_B^r$ of the utility for using a bicycle for group $r \in R^c, c \in IC$ comprises two parts: the monetary cost $OM_B^r$ which reflects the bicycle's depreciation cost, and the travel time-related costs $VOT_B^c T_B^r$. Since bicycle travel times $T_B^r$ are less affected by congestion due to the flexibility of bicycles, it is treated as a parameter rather than a variable. Therefore, the utility $V_B^r$ for using a bicycle can be represented by Equations (5.3).

$$V_{AV}^r = -\beta_0^c (OM_{AV}^r + VOT_{AV}^c T_{AV}^r) - \beta_1 (1-\alpha), \qquad \forall r \in R^c, c \in IC \tag{5.2}$$

$$V_B^r = -\beta_0^c (OM_B^r + VOT_B^c T_B^r), \qquad \forall r \in R^c, c \in IC \tag{5.3}$$

**Demand calculation:**

Knowing the probability of using SAVs, we can determine the demand $D^r_{AV}$ for SAVs for group $r \in R^c$, $c \in IC$ using Constraints (5.4), which is the total number of trips $n^r$ in group $r \in R^c, c \in IC$ multiplied by their probability $P^r_{AV}$ of choosing SAVs. The addition and subtraction of 0.5 on either side of Constraints (5.4) ensures that the demand will take the integer value closest to the value of $n^r P^r_{AV}$.

$$n^r P^r_{AV} - 0.5 < D^r_{AV} \leq n^r P^r_{AV} + 0.5, \qquad \forall r \in R^c, c \in IC. \tag{5.4}$$

### 5.3.4 SAV service planning and operation modelling

In this subsection, we first present the modelling of pricing strategies. We then introduce the mathematical models for fleet management (fleet sizing, initial fleet distribution and service quality) and operations (trip assignment and vehicle routing). The notations used in this subsection are summarised in Table 5.3.2.

*Table 5.3.2: Notation used for SAV service planning and operation modelling.*

| Notation | Description |
|---|---|
| Parameters | |
| $co$ | Unit driving operational cost of an SAV, in euros/km. |
| $cd$ | Penalty for drop-off delay of passengers, in euros/time step. |
| $cf$ | Depreciation cost in one hour for using an SAV, in euros/vehicle. |
| $\mathscr{M}^r$ | Big M-parameter, where $r \in R^c, c \in IC$. |
| $\mathscr{M}_{ijt_1t_2}$ | Big M-parameter, where $(i,j) \in L, t_1, t_2 \in T$, if $t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij}$. |
| Variables | |
| $p^0$ | Initial base fare for using an SAV, in euros/trip. |
| $p$ | Travel distance-related price for using an SAV, in euros/km. |
| $p^c$ | Travel distance-related price for using an SAV for income class $c \in IC$, in euros/km. |
| $S^r$ | Total number of trips served by SAVs from group $r$, where $r \in R^c, c \in IC$. |
| $PF^r_{i_{t_1} j_{t_2}}$ | Passenger flow in group of trips $r \in R^c, c \in IC$ served by an SAV on road link $(i,j)$, from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G, a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| $O$ | SAV fleet size. |
| $O_i$ | Initial distribution of SAVs at parking node $i \in N_P$ at the beginning of a day. |
| $E^r_t$ | Total number of passengers in group of trips $r \in R^c, c \in IC$ arriving at time $t \in T$. |
| $F_{i_{t_1} j_{t_2}}$ | Vehicle flow in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |
| $W_{i_t}$ | Total number of vehicles parking at node $i \in N_P$ from time instant $t$ to $t+1$, with $t \in T \setminus \{\mathscr{T}\}$. |
| $Z^r_t$ | Binary variable which is 1 when Constraint (5.23) is active, and 0 otherwise. |
| $X_{i_{t_1} j_{t_2}}$ | Binary variable which is 1 when any vehicle travels in road link $(i,j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. |
| $A^r_t$ | Binary variable which is 1 when at least one trip in group $r \in R^c, c \in IC$ arrives at time $t \in T$, and 0 otherwise. |

**Pricing strategy**

From the perspective of the SAV service provider, we aim to explore three distinct pricing strategies: (1) charging travellers an initial base fare followed by a distance-related fare; (2) charging travellers solely based on their travel distance; (3) implementing a pricing strategy that varies according to travellers' income class level, which is an income class-based, distance-related pricing scheme. The detailed explanation and equations for these three pricing strategies are presented below.

**Pricing strategy 1. Base fare + distance-based fare**

This pricing strategy is among the most commonly employed. The base fare $p^0$, a fixed amount charged at the ride's outset, ensures drivers earn a minimum amount per trip from the moment they accept a ride. This is crucial for short trips, where time and distance fares may not sufficiently compensate the driver. Additionally, it may deter the use of SAVs for short distances. Introducing a base fare, as discussed in this chapter, can improve bicycle use, offering a more eco-friendly transportation alternative. Following the base fare $p^0$, a distance-based fare $p$ is introduced based on the shortest travel distance $sd^r$. The out-of-pocket money $OM^r_{AV}$ for travellers in group $r \in R^c$ belong to class $c \in IC$ is obtained following Equation (5.5).

$$OM^r_{AV} = p^0 + sd^r p, \qquad \forall r \in R^c, c \in IC \tag{5.5}$$

**Pricing strategy 2. Distance-based fare**

Distance-based pricing is commonly used in some ride-sharing platforms, delivery services, and some traditional taxi services. It offers a straightforward price structure, which is favoured in markets where consumers prefer predictable pricing structures. The out-of-pocket money paid by travellers will be obtained using a distance-based fare $p$ times the shortest travel distance $sd^r$, as shown in Equation (5.6).

$$OM^r_{AV} = sd^r p, \qquad \forall r \in R^c, c \in IC \tag{5.6}$$

**Pricing strategy 3. Income class-based fare**

Income class-based pricing adjusts fares based on the income level of travellers. For simplicity, we focus solely on income class-based distance-based pricing strategies. The out-of-pocket money for travellers in group $r \in R^c, c \in IC$ is expressed in Equation (5.7), where $p^c$ is a distance-based fare for income class $c \in IC$.

$$OM^r_{AV} = sd^r p^c, \qquad \forall r \in R^c, c \in IC \tag{5.7}$$

**Fleet management and operation**

This section outlines the constraints that describe detailed SAV operations, including trip assignment, routing, and parking. Based on these operations and the endogenous SAV demand, service providers can make the most rational fleet management decisions. These include determining fleet size, setting the initial distribution of the fleet across parking depots, and ensuring service quality. The constraints are detailed as follows.

**Service rate calculation:**

SAV operators can reject those requests that bring no profits to the company. Under this situation, the number of served trips $S^r$ in group $r \in R^c, c \in IC$ can be less than or equal to the group's demand $D^r_{AV}$ for SAV services, which is ensured by Constraints (5.8). Constraint (5.9) determines the endogenously determined service rate $\alpha$ of the SAV service as the total number of served trips divided by the total demand for SAV. This is a non-linear constraint because of the product of continuous variable $\alpha$ and integer variables $D^r_{AV}$.

$$S^r \leq D^r_{AV}, \qquad \forall r \in R^c, c \in IC \tag{5.8}$$

$$\alpha \sum_{r \in R^c, c \in IC} D^r_{AV} = \sum_{r \in R^c, c \in IC} S^r \tag{5.9}$$

**Trip assignment and vehicle routing constraints:**

Constraints (5.10) ensure that passengers in group $r \in R^c, c \in IC$ are picked up by the SAVs at the origin node $o^r$ at departure time $a^r$. After departure, SAVs are allowed to choose different paths to avoid traffic congestion, even though they serve passengers with the same income class and in the same group. To capture the different arrival times of the SAVs, Constraints (5.11) and (5.12) are defined to ensure that the number of served trips $S^r$ in group $r \in R^c, c \in IC$ is equal to the number of trips arriving at destination $d^r$ at different times. Constraints (5.13) and (5.14) ensure that the passengers are left at the destination node $d^r$ upon the SAV's first arrival, and there will be no return trips to the origin node $o^r$ by the SAVs after they depart. Constraints (5.15) make sure that the vehicle flow circulation $F_{i_{t_1} j_{t_2}}$ on the time-space network must be greater than or equal to the passenger flow $PF^r_{i_{t_1} j_{t_2}}$. The vehicle flows describe not only the delivery process of a passenger but also the empty relocation process.

$$S^r = \sum_{j_t | \left(o^r_{a^r}, j_t\right) \in G} PF^r_{o^r_{a^r} j_t}, \qquad \forall r \in R^c, c \in IC \tag{5.10}$$

$$S^r = \sum_{t \in T | a^r + st^r \leq t \leq b^r} E^r_t, \qquad \forall r \in R^c, c \in IC \tag{5.11}$$

$$E_t^r = \sum_{i_{t_1} | \left(i_{t_1}, d_t^r\right) \in G} PF^r_{i_{t_1} d_t^r}, \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.12}$$

$$\sum_{j_{t_1} | \left(d_t^r, j_{t_1}\right) \in G} PF^r_{d_t^r j_{t_1}} = 0, \qquad \forall r \in R^c, c \in IC, a^r \leq t \leq b^r \tag{5.13}$$

$$\sum_{i_{t_1} | \left(i_{t_1}, o_t^r\right) \in G} PF^r_{i_{t_1} o_t^r} = 0, \qquad \forall r \in R, c \in IC, a^r \leq t \leq b^r \tag{5.14}$$

$$\sum_{r \in R, c \in IC} PF^r_{i_{t_1} j_{t_2}} \leq F_{i_{t_1} j_{t_2}}, \qquad \forall \left(i_{t_1}, j_{t_2}\right) \in G \tag{5.15}$$

**Flow conservation constraints:**

Constraints (5.16)–(5.18) describe the passenger flow conservation rule at node $i \in N$ in the network and the vehicle flow conservation at both normal and parking nodes.

$$\sum_{j_{t_1} | \left(j_{t_1}, i_t\right) \in G} PF^r_{j_{t_1} i_t} = \sum_{j_{t_2} | \left(i_t, j_{t_2}\right) \in G} PF^r_{i_t j_{t_2}}, \qquad \forall r \in R^c, c \in IC, a^r < t < b^r, \; i \in N, i \neq o^r, i \neq d^r \tag{5.16}$$

$$\sum_{j_{t_1} | \left(j_{t_1}, i_t\right) \in G, t_1 < t} F_{j_{t_1} i_t} = \sum_{j_{t_2} | \left(i_t, j_{t_2}\right) \in G, t < t_2} F_{i_t j_{t_2}}, \qquad \forall i \in N \setminus N_P, 0 < t < \mathscr{T} \tag{5.17}$$

$$\sum_{j_{t_1} | \left(j_{t_1}, i_t\right) \in G, t_1 < t} F_{j_{t_1} i_t} + W_{i_{t-1}} = \sum_{j_{t_2} | \left(i_t, j_{t_2}\right) \in G, t < t_2} F_{i_t j_{t_2}} + W_{i_t}, \qquad \forall i \in N_P, 0 < t < \mathscr{T} \tag{5.18}$$

**Fleet sizing and distribution constraints:**

Constraints (5.19) describe the distribution $O_i$ of SAVs across parking nodes $i \in N_P$ at the beginning of the studied period. When the optimisation period starts, these SAVs either set out from their parking locations to pick up passengers (indicated by $F_{i_0 j_t}$) or stay at the parking node (indicated by $W_{i_0}$), ready to receive orders from the SAV operator. The total SAV fleet size $O$ is expressed as the sum of the initial fleet distribution $O_i$ at each parking node $i \in N_P$, which is ensured by Constraint (5.20).

$$\sum_{j_t | (i_0, j_t) \in G} F_{i_0 j_t} + W_{i_0} = O_i, \qquad \forall i \in N_P \tag{5.19}$$

$$\sum_{i \in N_P} O_i = O \tag{5.20}$$

**Longest travel time determination:**

When determining the utility of using SAVs for group $r \in R^c, c \in IC$, we consider the longest travel time $T^r_{AV}$ experienced by all the travellers in this group. To obtain this value, we express the arrival times (indicated by binary variable $A^r_t$) of all travellers by Constraints (5.21). The longest travel time is then determined using Constraints (5.22)-(5.25), with the bounds specified by Constraints (5.24). In Constraints (5.23), $\mathscr{M}$ is a sufficiently large number. For each group of trips $r \in R^c, c \in IC$, there is only one longest travel time. This uniqueness is ensured by Constraints (5.25).

$$\frac{E^r_t}{n^r} \leq A^r_t \leq E^r_t, \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.21}$$

$$T^r_{AV} \geq A^r_t (t - a^r), \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.22}$$

$$T^r_{AV} \leq A^r_t (t - a^r) + \mathscr{M}(1 - Z^r_t), \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.23}$$

$$st^r \leq T^r_{AV} \leq b^r - a^r, \qquad \forall r \in R^c, c \in IC \tag{5.24}$$

$$\sum_{t | a^r + st^r \leq t \leq b^r} Z^r_t = 1, \qquad \forall r \in R^c, c \in IC \tag{5.25}$$

## 5.3.5 Traffic congestion modelling

Traffic congestion is included in this model by introducing time-dependent link capacity, which is the so-called spatial link capacity introduced in Van Essen & Correia (2019). Instead of including the traditional non-linear BPR function (Dafermos & Sparrow, 1969) directly into the model to determine the travel time, they define spatial link capacity which is the maximum number of vehicles that can pass road link $(i, j)$ from time instant $t_1$ to time instant $t_2$, denoted by $C_{i_{t_1} j_{t_2}} = (t_2 - t_1) Q_{ij} \left( \frac{1}{a} \left( \frac{t_2 - t_1}{t^{\min}_{ij}} - 1 \right) \right)^{\frac{1}{b}}$. Instead of determining the travel time directly, we select one option from multiple link-time-capacity combinations, as described by Constraints (5.26) and (5.27). Constraints (5.28) describe the first-in-first-out (FIFO) rule, meaning that the vehicles that enter the road link first will leave the road link first.

**Capacity constraints**

$$\sum_{t_1 | (i_t, j_{t_1}) \in G} X_{i_t j_{t_1}} \leq 1, \qquad \forall (i, j) \in L, t \in T \tag{5.26}$$

$$F_{i_{t_1} j_{t_2}} \leq \left\lfloor C_{i_{t_1} j_{t_2}} \right\rfloor X_{i_{t_1} j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{5.27}$$

**FIFO constraints**

$$t_1 + \sum_{t \in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \sum_{t \in T} X_{i_{t_2} j_t} (t - t_2) + \mathscr{M} \left( 1 - \sum_{t \in T} X_{i_{t_2} j_t} \right), \quad \forall (i,j) \in L, t_1 < t_2 \leq t_1 + t_{ij}^{\max} - t_{ij}^{\min} \tag{5.28}$$

### 5.3.6 Objective function

To maximise the SAV operator's profit, we establish an objective function (see Equation (5.29)) that accounts for the service's total income and its operational expenses. The total income is represented by the sum of the fares paid by all travellers, which is the first term of the objective function. It is crucial to highlight that this sum includes non-linear terms, which are the multiplications of continuous variables $OM_{AV}^r$ and integer variables $S^r$. Costs are divided into: (1) fleet depreciation, which is the purchase cost per vehicle divided by its lifespan given by $cf$, multiplied by the fleet size, and is represented by the second term in Equation (5.29); (2) operational costs, including fuel, maintenance, and insurance, determined by multiplying the total distance travelled by all SAVs by the cost per kilometre $co$, as indicated by the third term in Equation (5.29); and (3) penalties for late drop-offs $cd$, charged per time unit ($cd$) for delays beyond the shortest anticipated travel time, shown by the final term in Equation (5.29).

$$\max \sum_{r \in R^c, c \in IC} OM_{AV}^r S^r - cf \cdot O - co \left( \sum_{(i_{t_1}, j_{t_2}) \in G} l_{ij} F_{i_{t_1} j_{t_2}} \right) - cd \sum_{r \in R^c, c \in IC} \left( \sum_{t \in T} tE_t^r - (a^r + st^r) S^r \right) \tag{5.29}$$

The proposed models for the three pricing strategies are summarised below and are denoted as **[M0-1]**, **[M0-2]**, and **[M0-3]**, respectively.

**Pricing strategy 1: base fare + distance-based fare** **[M0-1]** Objective function (5.29), subject to Constraints (5.1)-(5.5) and (5.8)-(5.28).

**Pricing strategy 2: distance-based fare** **[M0-2]** Objective function (5.29), subject to Constraints (5.1)-(5.4), (5.6) and (5.8)-(5.28).

**Pricing strategy 3: income class-based fare** **[M0-3]** Objective function (5.29), subject to Constraints (5.1)-(5.4) and (5.7)-(5.28).

## 5.4 Solution methods

The model presented in Section 5.3 is a MINLP model, which is challenging to solve. This nonlinearity arises from three aspects: the nonlinear binary logit model described in Equation (5.1); the nonlinear constraint (5.9) for calculating the service rate, which involves the multiplication of a continuous variable with integer variables; and the nonlinear objective function (5.29) that includes products of continuous and integer variables.

This chapter proposes three distinct solution strategies to address the aforementioned nonlinearities effectively: (1) reformulating the model into a Mixed-integer Linear Programming (MILP) model by applying linearisation techniques, introduced in Section 5.4.1; (2) combining Particle Swarm Optimisation (PSO) with a reformulated MILP in an iterative framework; and (3) combining Bayesian Optimisation (BO) with a reformulated MILP.

### 5.4.1 Reformulated MILP model (M1)

Fan et al. (2023) propose a solution method to tackle the non-linearities brought by the binary logit model and the non-linear constraints. They transform the binary logit model into equations containing logarithmic terms and approximate these terms using an outer-inner approximation technique. Furthermore, they employ binary variables and the big-M method to reformulate the products of a continuous variable and integer variables. In this chapter, we employ these techniques to linearise Equations (5.1) and Constraints (5.9), as outlined by Constraints (5.81)-(5.103) in Appendix 5.A. We further linearise the nonlinear terms in the objective function, which involves the product of continuous and integer variables. The continuous variables represent pricing, while the integer variables represent the number of trips served. We now describe the linearisation of the objective function for the three proposed pricing strategies, followed by the presentation of the overall mathematical formulation for this method.

#### Objective function linearisation with different pricing strategies

The nonlinear objective function is shown in Equation (5.29) with three different pricing strategies described by Equation (5.5), (5.6), and (5.7). To linearise the first term

in the objective function, we firstly introduce binary variables $\overline{S}_h^r$ to discretise the integer variable $S^r$ for each group $r \in R^c, c \in IC$, as shown in Constraints (5.30). Here, parameter $\mathscr{H}^r = \lfloor \log_2(n^r) \rfloor$, which ensures that Constraints (5.30) still hold when all the trips $n^r$ in group $r \in R^c, c \in IC$ would use SAVs.

$$S^r = \sum_{h=0}^{\mathscr{H}^r} 2^h \overline{S}_h^r, \qquad \forall r \in R^c, c \in IC \tag{5.30}$$

Then, the non-linear term $\sum_{r \in R^c, c \in IC} OM_{AV}^r S^r$ can be reformulated as $\sum_{r \in R^c, c \in IC} \sum_{h=0}^{\mathscr{H}^r} 2^h OM_{AV}^r \overline{S}_h^r$. We additionally introduce continuous variables $\overline{Y}_h^r$ to replace $OM_{AV}^r \overline{S}_h^r$. Subsequently, we introduce the following constraints:

$$\overline{Y}_h^r \leq p_{\max}^r \overline{S}_h^r, \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\} \tag{5.31}$$

$$\overline{Y}_h^r \geq OM_{AV}^r - p_{\max}^r (1 - \overline{S}_h^r), \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\} \tag{5.32}$$

$$\overline{Y}_h^r \geq 0, \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\} \tag{5.33}$$

$$\overline{Y}_h^r \leq OM_{AV}^r, \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\}. \tag{5.34}$$

Here, $p_{\max}^r$ represents the upper-bound value of the variable $OM_{AV}^r$, where $r \in R, c \in IC$. The objective function is then updated as follows:

$$\begin{aligned} \max \sum_{r \in R, c \in IC} \sum_{h=0}^{\mathscr{H}^r} 2^h \overline{Y}_h^r - cf \cdot O - co \left( \sum_{(i_{t_1}, j_{t_2}) \in G} l_{ij} F_{i_{t_1} j_{t_2}} \right) \\ - cd \sum_{r \in R^c, c \in IC} \left( \sum_{t \in T} t E_t^r - (a^r + st^r) S^r \right). \end{aligned} \tag{5.35}$$

We now describe the process for appropriately determining the value of $p_{\max}^r$, which is the upper bound of the variable $OM_{AV}^r$. The variable $OM_{AV}^r$ denotes the monetary costs for using SAV service for travellers in group $r \in R^c, c \in IC$. The monetary costs associated with using SAVs must be kept within a reasonable range to ensure the service remains attractive to users. If these costs are set too high, the SAV service may become unaffordable, leading all users to opt for bicycles instead. Our objective, therefore, is to identify the maximum value of $OM_{AV}^r$ that still sustains at least one trip of demand for SAVs across at least one group of trips. In other words, the price should be high enough to optimise revenue without causing a complete loss of demand for SAV services within the entire system.

According to Equation (5.4), the demand $D_{AV}^r$ becomes zero if $n^r P_{AV}^r < 0.5$. To maintain a non-zero demand for SAVs, $P_{AV}^r$ must be at least $\frac{1}{2n^r}$. To ensure a non-zero demand for at least one group $r$, there must be at least one group that satisfies

Constraints (5.36). In these constraints, the left-hand side represents the probability of using SAVs for group $r$, which is derived from Equations (5.1), (5.2), and (5.3).

$$\frac{e^{-\beta_0^c(OM_{AV}^r+VOT_{AV}^c T_{AV}^r)-\beta_1(1-\alpha)}}{e^{-\beta_0^c(OM_{AV}^r+VOT_{AV}^c T_{AV}^r)-\beta_1(1-\alpha)}+e^{V_B^r}} \geq \frac{1}{2n^r}, \qquad \exists r \in R^c, c \in IC \tag{5.36}$$

By rewriting these constraints, we obtain the following:

$$OM_{AV}^r \leq \frac{-\beta_1(1-\alpha)-V_B^r+\ln(2n^r-1)}{\beta_0^c} - VOT_{AV}^c T_{AV}^r, \qquad \exists r \in R^c, c \in IC \tag{5.37}$$

The right-hand side of (5.37) is the largest when $\alpha = 1$ and $T_{AV}^r$ equals the minimum travel time $st^r$ at free-flow speed for group $r \in R^c, c \in IC$. Furthermore, the "there exist" symbol ($\exists$) can be reformulated as a "for all" symbol ($\forall$) by taking the maximum value of the right-hand side of Constraints (5.37) over all groups $r \in R^c, c \in IC$. This leads to Constraints (5.38).

$$OM_{AV}^r \leq \max_{r \in R^c, c \in IC} \left( \frac{-V_B^r+\ln(2n^r-1)}{\beta_0^c} - VOT_{AV}^c st^r \right), \qquad \forall r \in R^c, c \in IC \tag{5.38}$$

Thus, $p_{\max}^r$, which describes the upper bound of the variable $OM_{AV}^r$, can take the value of the right-hand side of Constraints (5.38). Noted that for each $r \in R^c, c \in IC$, $p_{\max}^r$ has the same value, allowing us to simplify $p_{\max}^r$ to $p_{\max}$. The value of $p_{\max}$ is provided in Equation (5.39). Consequently, Constraints (5.31) and (5.32) are updated to Constraints (5.40) and (5.41), respectively.

$$p_{\max} = \max_{r \in R^c, c \in IC} \left( \frac{-V_B^r+\ln(2n^r-1)}{\beta_0^c} - VOT_{AV}^c st^r \right) \tag{5.39}$$

$$\overline{Y}_h^r \leq p_{\max} \overline{S}_h^r, \qquad \forall r \in R^c, c \in IC, h \in \{0,1,\ldots,\mathscr{H}^r\} \tag{5.40}$$

$$\overline{Y}_h^r \geq OM_{AV}^r - p_{\max}(1-\overline{S}_h^r), \qquad \forall r \in R^c, c \in IC, h \in \{0,1,\ldots,\mathscr{H}^r\} \tag{5.41}$$

The overall mathematical formulations for the three proposed pricing strategies are summarised as follows:

**Pricing strategy 1: base fare + distance-based fare** **[M1-1]** Objective function (5.35) with Constraints (5.2)-(5.5), (5.8)-(5.28), (5.30), (5.33), (5.34), (5.38)-(5.41), the linearisation-related constraints (5.81)-(5.103), and the non-negativity constraints (5.104), (5.105), (5.111)-(5.130) in Appendix 5.A.

**Pricing strategy 2: distance-based fare** **[M1-2]** Objective function (5.35), subject to Constraints (5.2)-(5.4), (5.6), (5.8)-(5.28), (5.30), (5.33), (5.34), (5.38)-(5.41), the linearisation-related constraints (5.81)-(5.103), and the non-negativity constraints (5.105), (5.111)-(5.130) in Appendix 5.A.

**Pricing strategy 3: income class-based fare** **[M1-3]** Objective function (5.35), subject to Constraints (5.2)-(5.4), (5.7), (5.8)-(5.28), (5.30), (5.33), (5.34), (5.38)-(5.41), the linearisation-related constraints (5.81)-(5.103), and the non-negativity constraints (5.106)-(5.130) in Appendix 5.A.

### 5.4.2 Particle Swarm Optimisation (PSO) embedded with an iterative process of solving a reformulated MILP model (M2)

Linearising the binary logit model with the outer-inner approximation method requires a significant number of binary variables and constraints. Specifically, the number of these variables and constraints expands with the growth in the number of groups, thereby escalating the computational load. In this section, we introduce a hybrid solution approach that integrates Particle Swarm Optimisation (PSO) with a reformulated MILP model, referred to as M2. In M2, the probability $P_{AV}^{r}$ of utilising SAV services for each group $r \in R^{c}, c \in IC$ is not treated as a variable, but as a fixed parameter provided as input. Solving M2 to optimality yields the optimal objective function value under these given probabilities. In PSO, we define the probabilities $P_{AV}^{r}$ for all groups $r \in R^{c}, c \in IC$ as a candidate solution, dubbed a particle. The primary objective of the PSO is to move particles around in the search space to find the best objective function value of M2.

We first present the framework of the proposed algorithm in Figure 5.4.1. The detailed formulation of the reformulated MILP model (M2) is introduced in a later section. Following this, Section 5.4.2 details the tailored PSO procedure specifically designed for our problem. The concepts and terms mentioned in Figure 5.4.1, such as particles, *pbest*, *gbest*, velocities, and positions, are also described in a later section.

#### Inner loop: Iterative process of solving a reformulated MILP model (M2)

In the reformulated MILP model (M2), the pricing variables, fleet sizing variable, initial fleet distribution variables, and service rate variable will be endogenous variables, while the probabilities of using SAVs for each group of trips will be considered exogenous variables. Treating probabilities as exogenous variables enables us to transform

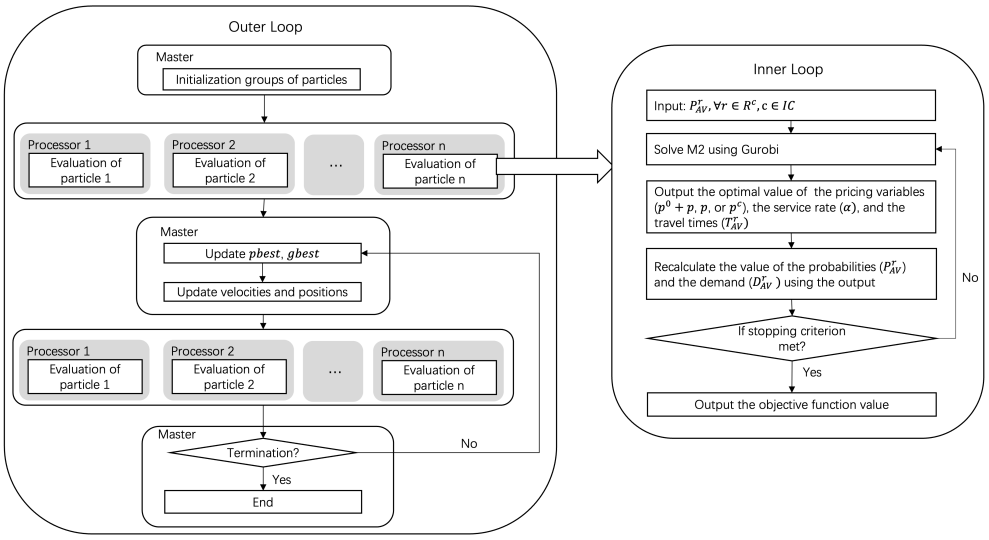



*Figure 5.4.1: Flowchart of the particle swarm optimisation framework embedded with an iterative process of solving a reformulated MILP model (M2).*

the non-linear binary logit model into linear constraints within the model, thus facilitating an easier solution process. First, rewriting Equation (5.1) and applying a logarithm give us the following equations:

$$V^r_{AV} = \ln P^r_{AV} - \ln(1 - P^r_{AV}) + V^r_B, \qquad \forall r \in R^c, c \in IC. \tag{5.42}$$

Combining Equation (5.42) with Equation (5.2) establishes a link between the probabilities $P^r_{AV}$ of using SAV, the price variables $p^0, p, p^c$, travel time $T^r_{AV}$, and service rate $\alpha$. However, when generating the exogenous variable $P^r_{AV}$ via PSO, random variations are introduced to these values. Directly replacing the non-linear binary logit model with Equation (5.42) may lead to infeasibility because of the integer nature of travel time $T^r_{AV}$, and the bounds on the pricing variables $p^0, p, p^c$ and service rate variable $\alpha$.

To maintain feasibility within the model, we relax Equations (5.42) by replacing the equality condition with an inequality, resulting in the relaxed constraints as shown in Constraints (5.43). This relaxation enables us to determine the minimum prices $p^0, p, p^c$ across all groups that satisfy these constraints.

$$V^r_{AV} \geq \ln P^r_{AV} - \ln(1 - P^r_{AV}) + V^r_B, \qquad \forall r \in R^c, c \in IC \tag{5.43}$$

Additionally, once $P^r_{AV}$ is predetermined as parameters before optimisation, Constraint (5.9) becomes linear, allowing demand variable $D^r_{AV}$ to be directly determined

using Equations (5.4). The mathematical formulations for **M2** under the three pricing strategies combine the following equations:

**Pricing strategy 1: base fare + distance-based fare** **[M2-1]** Objective function (5.35), subject to Constraints (5.2)-(5.5), (5.8)-(5.28), (5.30), (5.33), (5.34), (5.38)-(5.41), (5.43), and the non-negativity constraints (5.104), (5.105), (5.109)-(5.112), (5.115)-(5.130) in Appendix 5.A.

**Pricing strategy 2: distance-based fare** **[M2-2]** Objective function (5.35), subject to Constraints (5.2)-(5.4), (5.6), (5.8)-(5.28), (5.30), (5.33), (5.34), (5.38)-(5.41), (5.43), and the non-negativity constraints (5.105), (5.109)-(5.112), (5.115)-(5.130) in Appendix 5.A.

**Pricing strategy 3: income class-based fare** **[M2-3]** Objective function (5.35), subject to Constraints (5.2)-(5.4), (5.7), (5.8)-(5.28), (5.30), (5.33), (5.34), (5.38)-(5.41), (5.43), and the non-negativity constraints (5.106), (5.109)-(5.112), (5.115)-(5.130) in Appendix 5.A.

After solving the model, we obtain optimal values for the price, service rate, and travel time variables. These values are used to update the probability of using SAVs in each group, through Equation (5.1). These updated probabilities then serve as new inputs for M2. This iterative process continues until specific stopping criteria are met, which are: (1) reaching the maximum number of iterations; or (2) the difference in the objective function values between two consecutive iterations being less than or equal to a predetermined threshold. Figure 5.4.1 illustrates the detailed iterative process.

### Outer loop: Particle swarm optimisation (PSO)

PSO is a bio-inspired computational algorithm designed to solve optimisation problems by simulating the social behaviours observed in natural swarms. In PSO, a population of candidate solutions, referred to as particles, explores the search space to find (near-)optimal solutions. Each particle in the swarm adjusts its trajectory based on two key reference points: its own best-known position *pbest*, and the global best position *gbest* discovered by any member of the swarm. This mechanism guides each particle toward its personal and collective best positions, thereby facilitating convergence within the swarm.

In this chapter, the position of a particle is represented by a probability vector, with each element representing the probability of using SAV service $P_{AV}^{r}$ for each group

$r \in R^c, c \in IC$. PSO aims to find the optimal probability vector that renders the best objective function value for our problem.

### Step 1: Initialisation

**Initialise particle swarm** We initiate the PSO process by generating an initial population with a fixed number of particles. To ensure a strong starting point for optimisation, we create twice this number of particles and apply an elitism strategy to select the top-performing particles as the initial population.

When creating the particles, we construct the initial probability vector $\{P_{AV}^1, P_{AV}^2, \ldots, P_{AV}^r\}$ using a specialised grid search approach, rather than randomly generating values for $P_{AV}^r$ for each group $r \in R^c, c \in IC$. Each probability $P_{AV}^r$ has a distinct lower and upper bound. The upper bounds are determined using the minimum price, shortest travel time, and a no-rejection policy ($\alpha = 1$), providing the travellers with the highest utility as the upper bound. The lower bounds are set to 0.

Starting at the lower bound, each component of the vector is uniformly increased by a step size, calculated as $(\text{upper bound} - \text{lower bound})/\text{population size}$, until it reaches the upper bound. All components of the vector are increased simultaneously from their lower bounds to their respective upper bounds. Random generation is less efficient in our problem due to Constraints (5.43). If even one $P_{AV}^r$ is randomly assigned an extremely high value, it could restrict the price variables to lower values. This approach may restrict the search to a limited feasible region, resulting in particles with performance that are too similar to one another.

**Initialise particles' velocity** Instead of using the traditional random generation for the initial velocity, we utilise the distance information among the initial population to generate their initial velocities. This approach is inspired by the differential evolution (DE) algorithm, which leverages the spatial relationships between particles to establish their initial momentum. This method is effective in suggesting a better starting direction. In this chapter, for each particle that does not possess the *gbest* fitness value, we set the initial velocity as the distance between its current location and the location of *gbest*. For a particle with the *gbest* value, we determine its initial velocity using its distance from a randomly chosen *pbest*.

**Initialise velocity clamping operator** Velocity clamps are a mechanism used to prevent the particles from moving too quickly or too slowly across the search space. If the velocity of a particle is too high, it may skip over good solutions or move out of the

feasible region. On the other hand, if the velocity is too low, the changes in probability vectors may fail to produce any meaningful variation in the solution (e.g., total demand remains unchanged).

We define $V^r_{\max}$ and $V^r_{\min}$ for $r \in R^c, c \in IC$ as velocity clamping parameters to control the maximum and minimum velocities a particle can attain. The maximum speed, $V^r_{\max}$, is defined as 20% of the difference between the upper and lower bounds of the probabilities for each $r \in R^c, c \in IC$, ensuring that the velocity of any particle remains within the interval $[-V^r_{\max}, V^r_{\max}]$.

The minimum speed, $V^r_{\min}$, for each group $r \in R^c, c \in IC$, is determined as $1/n^r$. This ensures that the SAV system has at least one additional demand or one less. However, it is not necessary to apply the minimum speed threshold to each group of trips $r \in R^c, c \in IC$, as long as there are changes in the total demand. When the total demand of trips remains unchanged, we randomly select some groups and adjust their velocity to either the minimum speed $V^r_{\min}$ or its negative counterpart, $-V^r_{\min}$, to ensure variations in the total demand.

**Step 2: Main loop (repeat until stopping criteria are met)**

**Step 2.1 Update velocities and positions** We employ two strategies to update velocities and positions: the first is a dynamic PSO method that incorporates cognitive and social components, which are weighted by random factors to induce stochastic behaviour. The second strategy, inspired by DE, focuses on the thorough exploitation of the solution space surrounding the best solution identified thus far.

Strategy 1: Dynamic PSO searching strategy

For each particle, we determine the new velocity using its current velocity $V$, the vector distance from its current location $X$ to its *pbest*, and the vector distance from its current location $X$ to the *gbest*, as described by the following velocity update formula:

$$V_{\text{new}} = w \cdot V + c_1(t) \cdot r_1 \cdot (pbest - X) + c_2(t) \cdot r_2 \cdot (gbest - X) \tag{5.44}$$

$$c_1(t) = \left(c_{1,\min} - c_{1,\max}\right) \frac{t}{t_{\max}} + c_{1,\max} \tag{5.45}$$

$$c_2(t) = \left(c_{2,\max} - c_{2,\min}\right) \frac{t}{t_{\max}} + c_{2,\min} \tag{5.46}$$

Here, $w$ represents the inertia weight. This value controls how much of the previous velocity of each particle is retained as it moves to the next

iteration. $c_1(t)$ and $c_2(t)$ represent the cognitive and social coefficients at iteration $t$, respectively. The cognitive coefficient, $c_1(t)$, influences learning from personal experience, while the social coefficient, $c_2(t)$, draws on the swarm's collective knowledge. These coefficients dynamically change over the course of the iterations, with $c_1(t)$ gradually decreasing and $c_2(t)$ gradually increasing. This reflects a strategic shift from the initial exploration, focusing on individual best solutions, to later stages of convergence towards the global best. The maximum and minimum values for these coefficients are denoted as $c_{1,\max}$ and $c_{1,\min}$ for $c_1(t)$, and $c_{2,\max}$ and $c_{2,\min}$ for $c_2(t)$, respectively. A sensitivity analysis is performed on these parameters, with the results presented in Appendix 5.B. Additionally, $t_{\max}$ indicates the maximum number of iterations before the algorithm stops. Random factors $r_1$ and $r_2$, which are values between 0 and 1, are also included to introduce stochastic elements into the update equations in each iteration. This approach ensures a balanced trade-off between exploration and exploitation throughout the search process.

Then, we apply the velocity clamping operator to ensure that each particle's velocity remains within the predefined range, while also ensuring that there are variations in the total demand.

Next, we update the position of each particle $X_{\text{new}}$ by adding the clamped velocity to its current position $X$:

$$X_{\text{new}} = X + V_{\text{new}} \tag{5.47}$$

After updating positions, we check if any particle has moved outside the search space. If so, we apply a bounding strategy to adjust the particle's position back within the search space limits.

Strategy 2: DE searching strategy

We adopt a strategy from DE to thoroughly exploit the feasible region around the *gbest* solution. This approach adjusts each particle's position towards the best solution found so far, enhancing the search for an optimal solution by integrating the relative positions of other particles in the solution space.

$$X_{\text{new}} = X_{gbest} + r_1(X_{gbest} - X) + r_2(X_1 - X_2) \tag{5.48}$$

Here, $X_1 - X_2$ represent the distance between two randomly selected particles. After updating the location, we check if the new location falls within

the probability boundaries. If it does not, we update the location with the closest boundary value.

**Step 2.2 Evaluate fitness** After obtaining the new locations of the particles, we evaluate the fitness of each particle by solving M2.

**Step 2.3 Update *pbest* and *gbest*** We then update the *pbest* for each particle. If a particle's fitness at its new location is superior to the fitness at its *pbest*, the *pbest* is updated to this new position. Simultaneously, we assess all updated *pbest* values and if any particle's *pbest* surpasses the current global best *gbest*, we update *gbest* accordingly.

**Step 3: Termination** Repeat the main loop outlined in Step 2 for a predetermined number of iterations or until a convergence criterion is met, such as when *gbest* shows negligible improvement over a set number of iterations. The final solution is represented by the position of *gbest*.

### 5.4.3 Parallel Bayesian Optimisation with a reformulated MILP model (M3)

Bayesian Optimisation (BO) has proven to be a powerful tool for exploring the parameter space efficiently (Bergstra et al., 2011; Liu et al., 2019; Swersky et al., 2013). It employs a surrogate model to guide the selection of the next sampling points, with the goal of identifying optimal parameters with minimal evaluations. Instead of resorting to binary variables and the big-M method to reformulate the products of continuous and integer variables in the objective function (5.29), BO provides a sequential search strategy that allows us to strategically target the best pricing and service rate parameters to maximise the objective function value. By treating the service rate and pricing variables as parameters in Constraint (5.9) and objective function (5.29), we reduce the number of binary variables and constraints in Model M1. Note that we still need to linearise the binary logit model using the outer-inner approximation method. The formulation of the reformulated MILP, referred to as M3, is detailed under the three pricing strategies as follows:

**Pricing strategy 1: base fare + distance-based fare [M3-1]** Objective function (5.29) with Constraints (5.2)-(5.5), (5.8)-(5.28), the linearisation-related Constraints (5.81)-(5.97), and the non-negativity constraints (5.111)-(5.130) in Appendix 5.A.

**Pricing strategy 2: distance-based fare** **[M3-2]** Objective function (5.29) with Constraints (5.2)-(5.4), (5.6), (5.8)-(5.28), the linearisation-related Constraints (5.81)-(5.97), and the non-negativity constraints (5.111)-(5.130) in Appendix 5.A.

**Pricing strategy 3: income class-based fare** **[M3-3]** Objective function (5.29) with Constraints (5.2)-(5.4), (5.7), (5.8)-(5.28), the linearisation-related Constraints (5.81)-(5.97), and the non-negativity constraints (5.111)-(5.130) in Appendix 5.A.

There are two main components in BO: the surrogate model and the acquisition function. The surrogate model is a probabilistic model (typically a Gaussian Process) that serves as a surrogate for the true objective function. As new data points are observed, this model is iteratively updated, thereby improving its accuracy and reducing uncertainty with each iteration. The acquisition function is used to determine the next sample point to evaluate by aiming to maximise the acquisition function's value. In this chapter, we do not introduce the details of these two components. Interested readers can refer to Liu et al. (2019) for a thorough explanation. The parameters that the BO targets under the three pricing strategies are $[p^0, p, \alpha]$, $[p, \alpha]$, and $[p^c, \alpha]$ with $c \in IC$, respectively. The bounds of $p^0$, $p$, $p^c$, and $\alpha$ are determined by combining Constraints (5.38) with Equations (5.5), (5.6), and (5.7). BO can only target the values within these ranges.

$$0 \leq p^0 \leq \max_{r \in R^c, c \in IC} \left( \frac{-V_B^r + \ln(2n^r - 1)}{\beta_0^c} - VOT_{AV}^c st^r \right) \tag{5.49}$$

$$0 \leq p \leq \max_{r \in R^c, c \in IC} \left( \frac{-V_B^r + \ln(2n^r - 1)}{\beta_0^c sd^r} - \frac{VOT_{AV}^c st^r}{sd^r} \right) \tag{5.50}$$

$$0 \leq p^c \leq \max_{r \in R^c, c \in IC} \left( \frac{-V_B^r + \ln(2n^r - 1)}{\beta_0^c sd^r} - \frac{VOT_{AV}^c st^r}{sd^r} \right), \qquad c \in IC \tag{5.51}$$

$$0 \leq \alpha \leq 1 \tag{5.52}$$

The framework of parallel BO with the reformulated MILP model (M3) is illustrated in Figure 5.4.2.

## 5.5 Case study of the city of Delft, in the Netherlands

We apply the proposed solution methods to a case study of Delft, in the Netherlands. This section first introduces the application setting in Section 5.5.1. We then test the methods on a small problem involving only 9 groups of trips to compare the performance of the three proposed methods in Section 5.5.2. Following this small case study,

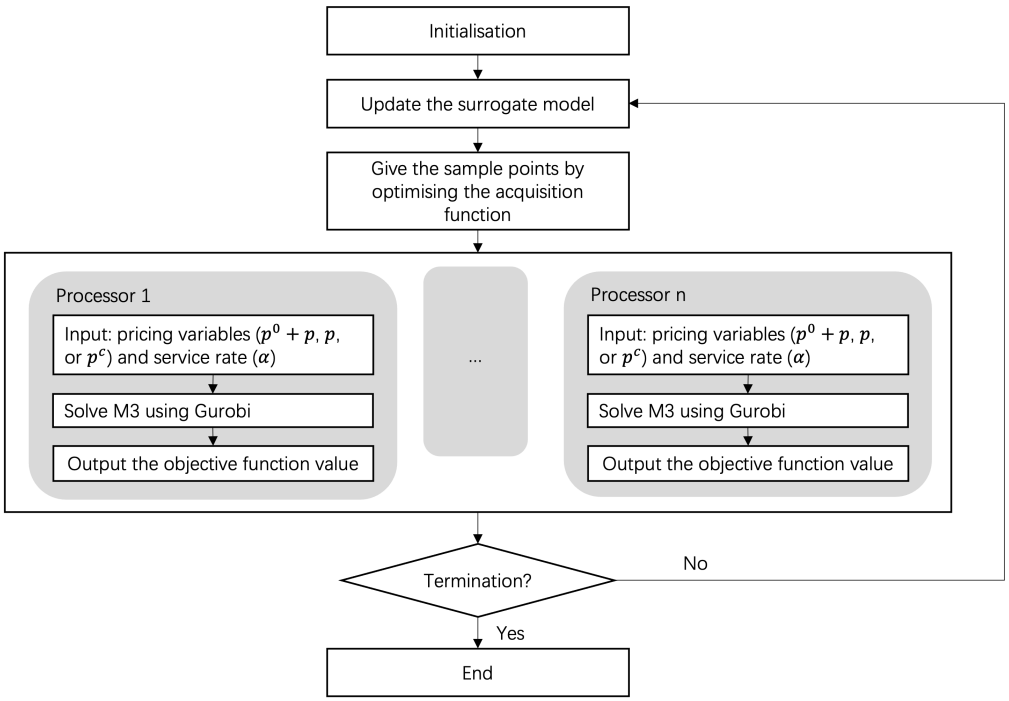



*Figure 5.4.2: Framework of the Bayesian optimisation with reformulated MILP model (M3).*

we extend the application to the entire city of Delft. The optimisation results are presented in Section 5.5.3.

### 5.5.1 Application setting

We introduce the application setting from the following perspectives: network description, mobility data, optimisation setting, parameter setting, and computational setting.

**Network description** We apply the proposed model and solution methods to a quasi-real case study of Delft, a city in the South-Holland province in the Netherlands, using the simplified road network detailed by Fan et al. (2023). The network consists of 35 nodes and 104 directed links that allow two-way traffic, as depicted in Figure 5.5.1. SAVs can navigate the entire network, but only seven nodes (3, 10, 11, 15, 19, 22, and 27) are designated as free parking depots, marked in red on the map. The road links are designed with capacities of 1600 vehicles for single lanes and 3200 for double lanes per hour, with speed limits of 50 km/h and 70 km/h, respectively, and a minimum speed of 5 km/h enforced on all roads.

**Mobility data** We utilised the Dutch mobility dataset (MON 2007/2008) (Correia & Van Arem, 2016; Liang et al., 2020; Fan et al., 2023) to extract detailed travel demand for bicycles, cars, and taxis between 7 am and 10 am. To represent these data effectively for the morning peak hour, we evenly distributed these trips over an hour, creating a total of 2933 trips. These were further aggregated into 45 groups based on similarity in trip characteristics.

**Optimisation setting** The optimisation considers a one-hour morning peak, divided into 24 time steps of 2.5 minutes each. To accommodate the trips from the parking depots, an additional 5 time steps are included before the peak, resulting in a total of 29 time steps for the complete optimisation period.

**Parameter setting** The parameters related to the network configuration, demand characteristics, and SAV operational dynamics in this case study are outlined here. The maximum travel time $t_{ij}^{\max}$ for SAVs to travel through link $(i, j) \in L$, is computed by dividing the length of each road link by the minimum travel speed of 5 km/h, ensuring that the vehicles adhere to speed limits. Conversely, the minimum travel time $t_{ij}^{\min}$ for SAVs to travel through link $(i, j) \in L$, is determined by dividing the road link length by the respective maximum speeds of 50 km/h or 70 km/h. Due to the time-space network

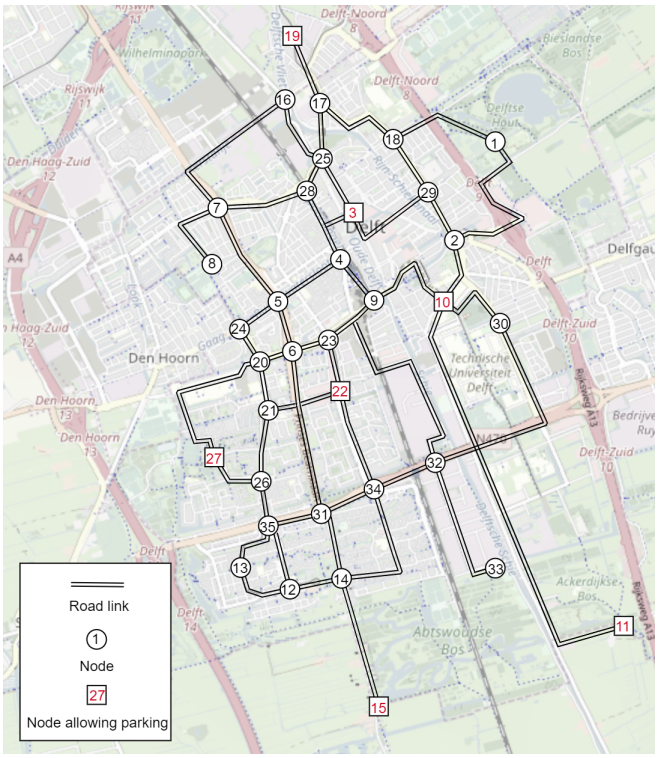



*Figure 5.5.1: Simplified road network of Delft used in the case study.*

model used in this study, the minimum travel time on each road link cannot fall below one time step. This ensures realism in the network model, as it prevents scenarios where a vehicle could traverse a link in zero time.

To obtain the shortest travel distances and times, $sd^r$ and $st^r$ for $r \in R^c, c \in IC$, we use the shortest path algorithm assuming vehicles can travel at free-flow speeds. The travel time for bicycles $T_B^r$ with $r \in R^c, c \in IC$ is determined by dividing the shortest path length by the average cycling speed of 12.4 km/h, reflective of typical Dutch cycling habits as noted by BicycleDutch (2018).

We set the logit model base parameter $\beta_0$ at 0.1 (Fan et al., 2023). The VOTT for using an SAV, $VOT_{AV}^c$ with $c \in IC$, is set at 6.6, 4.6, and 3.8 euros per hour for high, middle, and low-income travellers, respectively, as estimated by Kolarova et al. (2019). Similarly, the VOTT for using a bicycle, $VOT_B^c$ with $c \in IC$, is set at 24.9, 17.3, and 14.1 euros per hour for the respective income groups, following the same source.

The operational cost of SAVs, $co$, is determined to be 0.32 euros per km, based on the methodology from Bösch et al. (2018), while the depreciation cost $cf$ is set at 1.2 euros per vehicle per hour, as identified in Fan et al. (2022). Additionally, the delay

penalty *cd* is fixed at 0.2 euros per minute, following Liang et al. (2020). Finally, the parameters $a$ and $b$ in the BPR function are set to 2 and 4, respectively, aligning with the values suggested by Van Essen & Correia (2019).

**Algorithm parameter setting** We first introduce the parameter setting for the PSO-based solution algorithm introduced in Section 5.4.2. The population size of PSO is set to 12, and the maximum number of iterations is established at 30. The initial 10 iterations employ the dynamic PSO searching strategy, while the subsequent 20 iterations switch to the DE searching strategy. The inertia, $w$, is set at 0.2. Additionally, the cognitive and social coefficients are configured with the following values: $c_{1,\max}$ and $c_{2,\max}$ at 0.8, and $c_{1,\min}$ and $c_{2,\min}$ at 0.1. The inner loop terminates when either the number of iterations reaches 5 or the relative difference in the objective function values between two consecutive iterations is less than or equal to 5%.

For the BO-based solution algorithm introduced in Section 5.4.3, we maintain the same population size and maximum number of iterations as in the PSO-based solution algorithm to ensure consistency. Specifically, the number of individuals evaluated in each iteration is set to 12, and the algorithm terminates upon reaching the predefined maximum of 30 iterations.

**Computational setting** The proposed algorithms were implemented in Python 3.7 and the proposed MILP models were solved using Gurobi 9.5.2 on DelftBlue Supercomputer (Delft High Performance Computing Centre , DHPC).

### 5.5.2 Small case study

Before applying the three proposed solution methods to the real-life, city-sized case study of Delft, we first compare their performance on a smaller case study that includes only 9 groups of trips. These groups are randomly selected from the Dutch mobility dataset (MON 2007/2008). By testing the three proposed solution algorithms on this smaller case study, we can ensure that all the involved MILP models can be solved to optimality (resulting in a MILP gap of 0 for all methods in Table 5.5.1). This approach facilitates a more effective comparison on the performance of the three proposed solution algorithms.

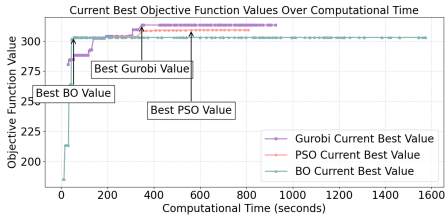
The optimisation results are presented in Table 5.5.1. In this table, we use Gurobi, PSO, and BO to represent the three proposed solution algorithms, as introduced in Sections 5.4.1, 5.4.2, and 5.4.3, respectively. We applied these algorithms to three different pricing strategies to assess their performance.

The objective function value is a key indicator of the quality of the found solutions, while computational time reflects the efficiency of these three algorithms. For PSO and BO, multiple MILP models are solved during the process. The table presents only the MILP gap and decision variable values that yield the best performance. As shown in Table 5.5.1, for all three pricing strategies, the proposed algorithms consistently find high-quality solutions within an acceptable time frame, demonstrating their performance. It is unsurprising to find that Gurobi provides the best objective function values in all the experiments. Following Gurobi, the PSO-based and BO-based methods also show similar objective function values. In terms of computational efficiency, the PSO-based method stands out as the most efficient, achieving good solutions in the shortest time among the three methods. Comparing the main decision variables—price, service rate, and fleet size—we find that these three algorithms yield similar optimal values for these variables, further proving the reliability of these methods.

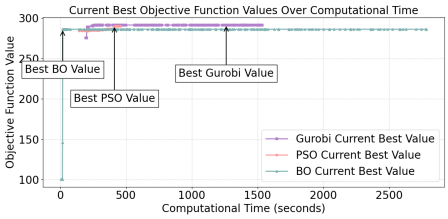
*Table 5.5.1: Comparison of optimisation results in the small case study using three proposed methods.*

| | Pricing strategy 1 | | | Pricing strategy 2 | | | Pricing strategy 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Gurobi | PSO | BO | Gurobi | PSO | BO | Gurobi | PSO | BO |
| Obj value | 313.27 | 310.55 | 302.99 | 291.43 | 290.28 | 286.03 | 319.69 | 312.24 | 309.78 |
| Computational time (s) | 1115 | 895 | 1575 | 1537 | 456 | 1587 | 4340 | 532 | 4196 |
| MILP gap (%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Price (€) | $p^0 = 4.93$<br>$p = 0.64$ | $p^0 = 4.89$<br>$p = 0.63$ | $p^0 = 4.44$<br>$p = 0.68$ | $p = 1.594$ | $p = 1.598$ | $p = 1.62$ | $p^{low} = 1.35$<br>$p^{mid} = 1.43$<br>$p^{high} = 1.76$ | $p^{low} = 1.29$<br>$p^{mid} = 1.48$<br>$p^{high} = 1.76$ | $p^{low} = 1.32$<br>$p^{mid} = 1.37$<br>$p^{high} = 1.79$ |
| Service rate (%) | 100 | 100 | 99 | 100 | 100 | 100 | 100 | 100 | 100 |
| Fleet size | 93 | 94 | 98 | 114 | 110 | 108 | 127 | 123 | 128 |

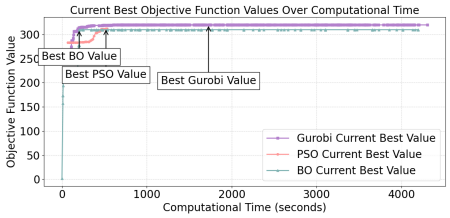
We further present the current best objective function values found by the three algorithms, along with the computational times, in Figure 5.5.2. Specifically, Figures 5.5.2a, 5.5.2b, and 5.5.2c show the performance comparisons under the pricing strategies 1, 2, and 3, respectively. These sub-figures demonstrate that the BO-based method converges to its best solution in the shortest amount of time compared to the other two algorithms. Initially, it spends some time exploring the feasible region, which leads to some solutions with lower objective function values. However, it quickly targets the most promising feasible region, achieving better results rapidly. For the PSO-based method, we display only the best solution found in each generation in these figures. Thanks to the elitism strategy employed during the initial population generation, the PSO-based method can already find a high-quality solution in the first generation.




*(a) Performance comparison under pricing strategy 1*




*(b) Performance comparison under pricing strategy 2*




*(c) Performance comparison under pricing strategy 3*

*Figure 5.5.2: Comparison of algorithm performance under the three pricing strategies in the small case study.*

## 5.5.3 Delft case study

In this section, we present the optimisation results of the Delft case study introduced in Section 5.5.1 and analyse the interplay among the endogenous decision variables. This analysis includes the impact of the three pricing strategies, congestion effects, and the spatial distribution of the rejection rate.

### Comparative analysis of algorithm performance

Due to the complexity and scale of the problem, it is challenging to solve all the MILP models to optimality. Therefore, we established specific stopping rules for different MILP models. For the first solution algorithm (denoted as Gurobi in Table 5.5.2), we set the time limit for solving the MILP model (M1) at 48 hours. For the second solution algorithm (denoted as PSO in Table 5.5.2), when solving the MILP model (M2), we first allow the solver to run for 20 minutes, during which it searches for the best possible solution. After these 20 minutes, we check the current optimality gap. If the gap is bigger than 3%, the solver continues. It stops when either the gap reaches 3% or the total runtime reaches 40 minutes. Similarly, for the third solution algorithm (denoted as BO in Table 5.5.2), we start by letting the solver run for 30 minutes to

search for the best possible solution. If, at the end of this period, the optimality gap is bigger than 3%, the solver continues. It stops when either the gap reaches 3% or the total runtime reaches 60 minutes. For comparison purposes, both the second and third solution algorithms are run to a 48-hour time limit. We believe that these times can be used in such a strategic context.

Table 5.5.2 presents the optimisation results of the Delft case study under the three pricing strategies. The results from Gurobi exhibit significant gaps when the 48-hour time limit is reached. For instance, with pricing strategy 1, the MILP gap reaches a value as high as 92.8%. The gaps under the second and third pricing strategies are lower, at 46.3% and 45.1%, respectively. Comparing the objective function values from the three solution algorithms, we observe that the PSO-based and BO-based algorithms outperform the Gurobi method with pricing strategies 1 and 2. With pricing strategy 3, the Gurobi method and the PSO-based algorithm exhibit very similar performance, with the Gurobi method slightly outperforming the PSO-based algorithm. However, the BO-based method achieves the best performance among the three methods.

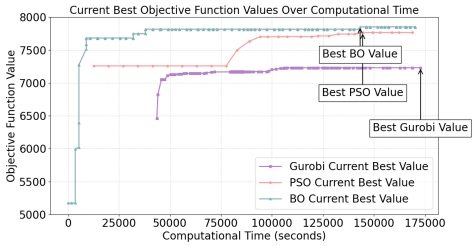
*Table 5.5.2: Comparison of optimisation results in the Delft case study using three proposed methods.*

| | Pricing strategy 1 | | | Pricing strategy 2 | | | Pricing strategy 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Gurobi | PSO | BO | Gurobi | PSO | BO | Gurobi | PSO | BO |
| Obj function value | 7230.91 | 7768.87 | 7852.19 | 7113.40 | 7207.06 | 7628.27 | 9143.37 | 9105.75 | 9341.43 |
| MILP gap (%) | 92.8 | 1.54 | 1.06 | 46.3 | 4.25 | 0.77 | 45.1 | 3.79 | 4.60 |
| Price (€) | $p^0 = 1.16$<br>$p = 2.54$ | $p^0 = 1.36$<br>$p = 1.10$ | $p^0 = 1.43$<br>$p = 1.11$ | $p = 1.09$ | $p = 1.16$ | $p = 1.32$ | $p^{low} = 0.99$<br>$p^{mid} = 1.17$<br>$p^{high} = 1.45$ | $p^{low} = 0.92$<br>$p^{mid} = 1.08$<br>$p^{high} = 1.58$ | $p^{low} = 0.94$<br>$p^{mid} = 1.09$<br>$p^{high} = 1.54$ |
| Service rate (%) | 100 | 98.71 | 100 | 99.78 | 95.23 | 98.26 | 95.72 | 94.74 | 90.85 |
| Fleet size | 596 | 964 | 926 | 1296 | 1088 | 855 | 1193 | 1266 | 1199 |

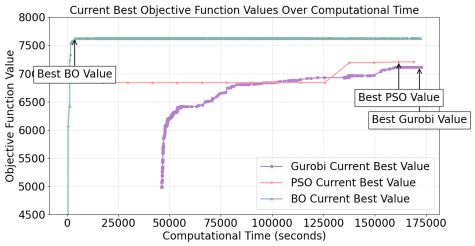
Figure 5.5.3 illustrates the trend of the best-found solutions over computational time up to the 48-hour time limit. From the figure, we can draw similar conclusions to those observed in the small case study: both the PSO-based and BO-based solution methods are capable of finding high-quality solutions in a shorter time compared to Gurobi.

Gurobi takes a long time to find the first feasible solution, and the quality of this solution is relatively poor compared to the other two methods. After finding the first feasible solution, it quickly improves, finding better solutions more rapidly. This observation suggests that providing a good initial solution could be beneficial for enhancing the performance of Gurobi. For the PSO-based approach, the first 10 iterations using the dynamic PSO strategy do not significantly improve the best solution. This is because the grid search approach and elitism strategy in the initial population generation

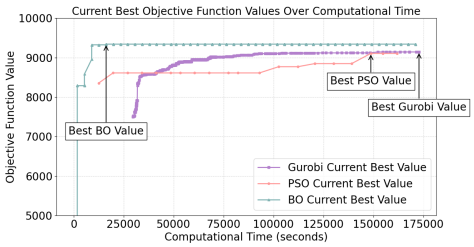
have already found a reasonably good starting solution. The BO-based method shows the same performance as in the small case study. It initially explores the feasible region with some random trials, leading to several solutions of lower quality, but then it efficiently targets the most promising feasible region.




*(a) Performance comparison under pricing strategy 1*




*(b) Performance comparison under pricing strategy 2*




*(c) Performance comparison under pricing strategy 3*

*Figure 5.5.3: Comparison of algorithm performance under three pricing strategies in the Delft case study.*

## Optimisation results

From the previous section, we observe the best performance using BO for the Delft case study. In this section, we analyse the optimisation results obtained using the BO-based method. The optimisation results are presented in Table 5.5.3.

We first investigate the optimal pricing strategy for SAV services and its impact on travellers' demand. The three proposed pricing strategies exhibit significantly different objective function values in the Delft case study. The third pricing strategy, income class-based pricing, yields the highest profit for the SAV service provider, as shown in Table 5.5.2. This strategy attracts more travellers to use the SAV service. The total demand under pricing strategy 3 is significantly higher than under pricing strategies 1 and 2.

*Table 5.5.3: Optimisation results of the Delft case study.*

| | Pricing strategy 1 | Pricing strategy 2 | Pricing strategy 3 |
|---|---|---|---|
| Total revenue (€) | 13898.91 | 12917.70 | 17626.79 |
| Total depreciation cost (€) | 1111.2 | 1026.0 | 1438.8 |
| Total operational cost(€) | 4254.52 | 3770.93 | 5663.06 |
| Total delay penalty (€) | 681.0 | 492.5 | 1183.5 |
| Total demand for SAVs (€) | 1263.0 | 1265.0 | 1792.0 |
| SAV demand share (%) | 43.06 | 43.13 | 61.10 |
| Percentage of satisfied demand (%) | 100 | 98.26 | 90.85 |
| Average price per trip (€) | 11 | 10.39 | 10.83 |
| Average delay per trip (min) | 2.7 | 1.98 | 3.63 |
| Average deliver time per trip | 16.30 | 14.60 | 17.43 |
| SAVs total travel distance (km) | 13295.38 | 11784.15 | 17697.06 |
| SAVs total deliver distance (km) | 11627.54 | 10266.68 | 15525.18 |
| SAVs total relocate distance (km) | 1667.84 | 1517.46 | 2171.87 |

In Table 5.5.2, we observe that high-income travellers are charged the highest unit distance-based rates, followed by middle-income and then low-income travellers. Higher-income travellers are less sensitive to price changes, making the higher rates still appealing for them to use the SAV service. A more detailed comparison can be seen in Figure 5.5.4. For pricing strategy 3, the demand difference between low, middle, and high-income classes is smaller compared to pricing strategies 1 and 2. The demand for SAVs among low and middle-income groups significantly increases under pricing strategy 3 compared to strategies 1 and 2. Interestingly, under pricing strategy 3, while the demand for low and middle-income class travellers increases, the demand for high-income class travellers decreases compared to pricing strategies 1 and 2. This is due to two factors: first, high-income travellers are charged higher fees. Second, high-income individuals with a higher value of travel time (VOTT) are more sensitive to increased travel time due to congestion effects.

Pricing strategy 1 imposes a base fare for every trip, meaning travellers must pay this fare regardless of distance. This strategy discourages the use of SAVs for short trips. As shown in Figure 5.5.4, for trips less than 5 km, pricing strategy 2 generates higher demand than pricing strategy 1 across all income classes. Pricing strategy 3 exhibits the same trend for low and middle-income travellers but not for high-income travellers, who are charged higher fees under this strategy. Thus, pricing strategy 1 is more sustainable as it encourages active modes of transport for short trips. Interestingly, the average trip revenue for pricing strategy 1 is the highest among the three strategies, as shown in Table 5.5.2. Despite having roughly the same demand as pricing strategy 2, pricing strategy 1 generates significantly more revenue for the SAV service provider.

Pricing strategy 2 considers only the distance-based fare. It attracts a similar num-

ber of travellers as pricing strategy 1. However, the SAV operator does not serve all trips, only the profitable ones, resulting in a smaller fleet, lower depreciation and operational costs, and reduced delay penalties.

Congestion effects are closely linked to fleet size, road link capacity, and the spatial distribution of demand. Pricing strategy 2 results in a less congested network with the fewest SAVs on the road, but it yields the lowest profit for the SAV service provider among the three strategies. The average delay per trip is similar for both pricing strategy 1 and pricing strategy 3. However, pricing strategy 3 generates the highest profits for the SAV service provider. From the provider's perspective, pricing strategy 3 is the best option. Although it slightly compromises delivery time due to congestion, it attracts more SAV users and generates higher profits.

Figure 5.5.4 illustrates the SAV demand for all income classes across trip lengths of 0-5 km, 5-10 km, and greater than 10 km. Several more conclusions can be drawn from the figures. Firstly, for short trips of less than 5 km, individuals from low and middle-income classes are particularly sensitive to the pricing strategy, and a fixed base fare is not appealing to them. Secondly, as trip length increases, the demand difference between low, middle, and high-income classes becomes more pronounced. Higher-income individuals are more likely to choose the SAV service due to the higher VOTT of using bicycles. Lastly, the willingness to use the SAV service increases with income, indicating that higher-income individuals are more inclined to opt for SAVs.


*(a)* *(b)* *(c)*

*Figure 5.5.4: SAV demand for all income classes with trip lengths of (a) 0-5 km, (b) 5-10 km, (c) >10 km.*

## 5.6 Conclusions and future research

This study addresses optimal pricing and fleet management for an SAV service provider in future urban mobility frameworks. In this setting, travellers choose between SAVs and bicycles, aligning with urban trends towards car-free environments. Travellers

request SAVs via apps, and the platform accepts or rejects them based on company benefits. Accepted requests prompt SAV dispatch to pick up customers. Our research proposes a Mixed-integer Nonlinear Programming (MINLP) model to explore optimal pricing strategies and assess varying traveller price sensitivities. Three different pricing strategies are investigated: base fare plus distance-based fare, distance-based fare only, and income class-based fare.

We develop three solution methods: first, we linearise the proposed model and solve the problem using a commercial solver; second, we propose a hybrid approach combining Particle Swarm Optimisation (PSO) with an iterative framework to solve a reformulated MILP model; third, we integrate Bayesian Optimisation (BO) with a reformulated MILP model to aid in complex objective function evaluations. We apply the proposed solution methods to a small case study and a case study of Delft, the Netherlands.

Conclusions can be drawn from both the algorithm performance perspective and the optimisation results. The three methods exhibit similar performance in the small case study, demonstrating their effectiveness. Among them, the PSO-based method is the most time-efficient. However, in the Delft case study, the BO-based method outperforms the other two methods. Regarding the optimisation results, the three proposed pricing strategies for SAV services have significantly different impacts on demand and profitability. Pricing strategy 1, which is the base fare and distance-related fare, tends to be the most environmentally sustainable, as it discourages the use of SAVs and promotes active transport for short trips ($< 5$ km). Pricing strategy 2 leads to less congestion and fewer SAVs on the road but results in the lowest profit. Pricing strategy 3, which is the income class-based fare, yields the highest profit for the SAV service provider by attracting a higher total demand compared to pricing strategies 1 and 2. High-income travellers are charged the highest rates, effectively leveraging their lower price sensitivity to increase overall revenue. Although pricing strategy 3 causes more congestion, it encourages low and middle-income users to utilise the SAV service, thereby enhancing social equality while being the most profitable for the SAV service provider. Travellers with different socio-demographic characteristics exhibit varying behaviours. Demand for SAV services increases with income, particularly for longer trips, with high-income individuals being more likely to choose SAVs.

Future research could explore several promising areas. Firstly, investigating dynamic pricing strategies that account for spatial-temporal congestion levels could provide more effective congestion management. Secondly, examining optimal ride-sharing pricing mechanisms could help alleviate congestion effects and increase fleet utilisation rates. Additionally, the pricing mechanisms of multiple SAV service providers could be studied to understand competitive dynamics and market impacts. Finally, de-

veloping heuristics to generate high-quality starting solutions for commercial solvers is a promising approach to ensure optimal outcomes. This is not covered in this research, as the solution from the PSO-based method may be infeasible for the reformulated MILP model (M1) due to their different feasible regions.

# Appendix

# 5.A Problem formulation

We summarise the complete problem formulation for the reformulated MILP model (M1) presented in Chapter 5, along with the notations for the sets, parameters, and variables below.

*Table 5.A.1: Notation for the sets, parameters, and variables*

| Notation | Description |
|---|---|
| **Set** | |
| $T$ | Set of time instants $T = \{0, 1, 2, \ldots, \mathscr{T}\}$ in the operational period. |
| $N$ | Set of physical nodes within the network. |
| $L$ | Set of road links connecting the nodes in set $N$. |
| $G$ | Set of links in the time-space network. |
| $N_P$ | Set of nodes that allowing parking for SAVs with $N_P \subseteq N$. |
| $IC$ | Set of income classes, categorised into low income $(low)$, middle income $(mid)$, and high income $(high)$. |
| $R^c$ | Set of groups of trips associated with a specific income class $c \in IC$. Each group $r \in R^c$ consists of trips that share the same characteristics, including origin, destination, departure time, latest arrival time, and income level $c$. |
| $M$ | Set of travel modes, consisting of automated vehicles $(AV)$ and bicycles $(B)$ as options. |
| $K$ | $= \{1, 2, \ldots, k, \ldots, \mathscr{K}\}$. Index set of predetermined breakpoints. |
| **Parameters** | |
| $\Delta t$ | Time step. |
| $l_{ij}$ | Length of road link $(i, j) \in L$. |
| $Q_{ij}$ | Capacity of road link $(i, j) \in L$ in number of vehicles per time step. |
| $t_{ij}^{\max}$ | Maximum travel time by car on road link $(i, j) \in L$. |
| $t_{ij}^{\min}$ | Minimum travel time by car on road link $(i, j) \in L$. |
| $C_{i_{t_1} j_{t_2}}$ | Spatial capacity of road link $(i, j) \in L$ in number of vehicles that fit on the road link from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. |
| $\alpha$ | Trip service rate when all the requests have to be accepted, %. |
| $o^r$ | Origin node for group of trips $r \in R^c, c \in IC$. |
| $d^r$ | Destination node for group of trips $r \in R^c, c \in IC$. |
| $a^r$ | Departure time for group of trips $r \in R^c, c \in IC$. |
| $b^r$ | Latest arrival time for group of trips $r \in R^c, c \in IC$. |
| $sd^r$ | Shortest travel distance for group of trips $r \in R^c, c \in IC$, in kilometres. |
| $st^r$ | Shortest travel time assuming free-flow speed for group of trips $r \in R^c, c \in IC$, in time steps. |
| $n^r$ | Total number of trips for group $r \in R^c, c \in IC$. |
| $V_B^r$ | Deterministic systematic component of the utility of bicycles for group of trips $r \in R^c, c \in IC$. |
| $OM_B^r$ | Monetary costs of travellers in group $r \in R^c, c \in IC$ using bicycles, in euros. |
| $\beta_0^c$ | Parameter converting generalised costs into utility for income class $c \in IC$, in utility/euro. |
| $\beta_1$ | Parameter converting service rate into utility. |

| | |
|---|---|
| $VOT_m^c$ | Travellers' value of travel time in class $c \in IC$ using mode $m \in M$, in euros/time step. |
| $T_B^r$ | Travel time of using bicycles for trips in group $r \in R^c, c \in IC$. |
| $co$ | Unit driving operational cost of an SAV, in euros/km. |
| $cd$ | Penalty for drop-off delay of passengers, in euros/time step. |
| $cf$ | Depreciation cost in one hour for using an SAV, in euros/vehicle . |
| $(u^k, \ln u^k)$ | Coordinates of the $k^{\text{th}}$ breakpoint. |

**Decision variables**

| | |
|---|---|
| $V_{AV}^r$ | Deterministic systematic component of travellers' utility for using an SAV in group $r \in R^c, c \in IC$. |
| $OM_{AV}^r$ | Monetary costs of travellers in group $r \in R^c, c \in IC$ using SAVs, in euros. |
| $T_{AV}^r$ | Longest SAVs travel time for group $r \in R^c, c \in IC$. |
| $P_{AV}^r$ | Probability to choose SAVs for the trips in group $r \in R^c, c \in IC$. |
| $D_{AV}^r$ | Total number of trips using SAVs in group $r \in R^c, c \in IC$. |
| $\alpha$ | Trip service rate. |
| $p^0$ | Initial base fare for using SAVs, in euros/trip. |
| $p$ | Travel distance-related price for using an SAV, in euros/km. |
| $p^c$ | Travel distance-related price for using an SAV for income class $c \in IC$, in euros/km. |
| $S^r$ | Total number of trips served by SAVs from group $r$, where $r \in R^c, c \in IC$. |
| $PF_{i_{t_1} j_{t_2}}^r$ | Passenger flow in the group of trips $r \in R^c, c \in IC$ served by an SAV in road link $(i, j)$, from time instant $t_1$ to $t_2$. Only defined for $(i_{t_1}, j_{t_2}) \in G, a^r \leq t_1 < t_2 \leq b^r$. If $t_1 = a^r$, then $i = o^r$. |
| $O$ | SAV fleet size. |
| $O_i$ | Initial distribution of SAVs at parking node $i \in N_P$ at the beginning of a day. |
| $E_t^r$ | Total number of passengers in group of trips $r \in R^c, c \in IC$ arriving at time $t \in T$. |
| $F_{i_{t_1} j_{t_2}}$ | Vehicle flow in road link $(i, j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$. Note that when $t_1 = 0$, $i \in N_P$, meaning that SAVs have to depart from the parking nodes at the beginning of a day. |
| $W_{i_t}$ | Total number of vehicles parking at node $i \in N_P$ from time instant $t$ to $t+1$, with $t \in T \setminus \{\mathscr{T}\}$. |
| $Z_t^r$ | Binary variable which is 1 when Constraint (5.23) is active, and 0 otherwise. |
| $X_{i_{t_1} j_{t_2}}$ | Binary variable which is 1 when any vehicle travels in road link $(i, j)$ from time instant $t_1$ to $t_2$, where $(i_{t_1}, j_{t_2}) \in G$, and 0 otherwise. |
| $A_t^r$ | Binary variable which is 1 when at least one trip in group $r \in R^c, c \in IC$ arrives at time $t \in T$, and 0 otherwise. |
| $LN_{AV}^r$ | Auxiliary continuous variable, where $r \in R^c, c \in IC$. |
| $LN_B^r$ | Auxiliary continuous variable, where $r \in R^c, c \in IC$. |
| $\lambda_r^k$ | Binary variable indicating whether an interval $[u^k, u^{k+1}]$ is active or not, where $k \in \{1, 2, \ldots, k, \ldots, \mathscr{K}-1\}$, $r \in R^c, c \in IC$. |
| $\theta_r^k$ | Convex combination coefficient for breakpoint $k \in K$ for group of trips $r \in R^c, c \in IC$. |
| $\overline{\lambda}_r^k$ | Binary variable indicating whether an interval $[1-u^{k+1}, 1-u^k]$ is active or not, where $k \in \{1, 2, \ldots, k, \ldots, \mathscr{K}-1\}$, $r \in R^c, c \in IC$. |
| $\overline{\theta}_r^k$ | Convex combination coefficient for breakpoint $k \in K$ for group of trips $r \in R^c, c \in IC$. |
| $\overline{D}_h$ | Binary variables utilised for discretising integer variables, where $h \in \{0, 1, \ldots, \mathscr{H}\}$. |
| $Y_h$ | Continuous variables utilised for describing the value of the integer variables, where $h \in \{0, 1, \ldots, \mathscr{H}\}$. |

| | |
|---|---|
| $\overline{Y}_h^r$ | Continuous variables utilised for describing the value of the product of integer variables and continuous variables, where $r \in R^c, c \in IC, h \in \{0,1,\ldots,\mathscr{H}\}$. |
| $\overline{S}_h^r$ | Binary variables utilised for discretising integer variables, where $r \in R^c, c \in IC, h \in \{0,1,\ldots,\mathscr{H}\}$. |

*Mixed integer linear program*

$$\begin{aligned}\max \sum_{r\in R, c\in IC} \sum_{h=0}^{\mathscr{H}^r} 2^h \overline{Y}_h^r - cf\cdot O - co\left(\sum_{(i_{t_1}, j_{t_2})\in G} l_{ij} F_{i_{t_1} j_{t_2}}\right) \\ - cd \sum_{r\in R^c, c\in IC}\left(\sum_{t\in T} tE_t^r - (a^r + st^r)S^r\right).\end{aligned} \tag{5.53}$$

subject to:

$$OM_{AV}^r = \begin{cases} p^0 + sd^r p & \text{if pricing strategy 1,} \\ sd^r p & \text{if pricing strategy 2,} \\ sd^r p^c & \text{if pricing strategy 3,} \end{cases} \qquad \forall r \in R^c, c \in IC \tag{5.54}$$

$$V_{AV}^r = -\beta_0^c (OM_{AV}^r + VOT_{AV}^c T_{AV}^r) - \beta_1(1-\alpha), \qquad \forall r \in R^c, c \in IC \tag{5.55}$$

$$n^r P_{AV}^r - 0.5 < D_{AV}^r \le n^r P_{AV}^r + 0.5, \qquad \forall r \in R^c, c \in IC \tag{5.56}$$

$$S^r \le D_{AV}^r, \qquad \forall r \in R^c, c \in IC \tag{5.57}$$

$$S^r = \sum_{j_t | (o_{a^r}^r, j_t)\in G} PF_{o_{a^r}^r j_t}^r, \qquad \forall r \in R^c, c \in IC \tag{5.58}$$

$$S^r = \sum_{t\in T | a^r + st^r \le t \le b^r} E_t^r, \qquad \forall r \in R^c, c \in IC \tag{5.59}$$

$$E_t^r = \sum_{i_{t_1} | (i_{t_1}, d_t^r)\in G} PF_{i_{t_1} d_t^r}^r, \qquad \forall r \in R^c, c \in IC, t \in T \tag{5.60}$$

$$\sum_{j_{t_1} | (d_t^r, j_{t_1})\in G} PF_{d_t^r j_{t_1}}^r = 0, \qquad \forall r \in R^c, c \in IC, a^r \le t \le b^r \tag{5.61}$$

$$\sum_{i_{t_1} | (i_{t_1}, o_t^r)\in G} PF_{i_{t_1} o_t^r}^r = 0, \qquad \forall r \in R^c, c \in IC, a^r \le t \le b^r \tag{5.62}$$

$$\sum_{j_{t_0}|(j_{t_0},i_{t_1})\in G} PF^r_{j_{t_0}i_{t_1}} = \sum_{j_{t_2}|(i_{t_1},j_{t_2})\in G} PF^r_{i_{t_1}j_{t_2}}, \forall r \in R^c, c \in IC, \quad a^r < t_1 < b^r, i \in N, i \neq o^r, i \neq d^r \tag{5.63}$$

$$\sum_{r\in R^c c\in IC} PF^r_{i_{t_1}j_{t_2}} \leq F_{i_{t_1}j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{5.64}$$

$$\sum_{j_{t_1}|(j_{t_1},i_t)\in G, t_1<t} F_{j_{t_1}i_t} = \sum_{j_{t_2}|(i_t,j_{t_2})\in G, t<t_2} F_{i_t j_{t_2}}, \forall i \in N \setminus N_P, 0 < t < \mathscr{T} \tag{5.65}$$

$$\sum_{j_{t_1}|(j_{t_1},i_t)\in G, t_1<t} F_{j_{t_1}i_t} + W_{i_{t-1}} = \sum_{j_{t_2}|(i_t,j_{t_2})\in G, t<t_2} F_{i_t j_{t_2}} + W_{i_t}, \qquad \forall i \in N_P, 0 < t < \mathscr{T} \tag{5.66}$$

$$\sum_{j_t|(i_0,j_t)\in G} F_{i_0 j_t} + W_{i_0} = O_i, \forall i \in N_P \tag{5.67}$$

$$\sum_{i\in N_P} O_i = O \tag{5.68}$$

$$\frac{E^r_t}{n^r} \leq A^r_t \leq E^r_t, \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.69}$$

$$T^r_{AV} \geq A^r_t (t - a^r), \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.70}$$

$$T^r_{AV} \leq A^r_t (t - a^r) + (b^r - a^r)(1 - Z^r_t), \qquad \forall r \in R^c, c \in IC, a^r + st^r \leq t \leq b^r \tag{5.71}$$

$$\sum_{t|a^r+st^r\leq t\leq b^r} Z^r_t = 1, \qquad \forall r \in R^c, c \in IC \tag{5.72}$$

$$\sum_{t_2|(i_{t_1},j_{t_2})\in G} X_{i_{t_1}j_{t_2}} \leq 1, \qquad \forall (i,j) \in L, t_1 \in T \tag{5.73}$$

$$F_{i_{t_1}j_{t_2}} \leq \left\lfloor C_{i_{t_1}j_{t_2}} \right\rfloor X_{i_{t_1}j_{t_2}}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{5.74}$$

$$t_1 + \sum_{t\in T} X_{i_{t_1}j_t}(t - t_1) \leq t_2 + \sum_{t\in T} X_{i_{t_2}j_t}(t - t_2) + (t_1 + t^{\max}_{ij} - t_2)\left(1 - \sum_{t\in T} X_{i_{t_2}j_t}\right), \quad \forall (i,j) \in L, t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij} \tag{5.75}$$

$$S^r = \sum_{h=0}^{\mathscr{H}^r} 2^h \overline{S}^r_h, \qquad \forall r \in R^c, c \in IC \tag{5.76}$$

$$\overline{Y}^r_h \leq p_{\max} \overline{S}^r_h, \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\} \tag{5.77}$$

$$\overline{Y}^r_h \geq OM^r_{AV} - p_{\max}(1 - \overline{S}^r_h), \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\} \tag{5.78}$$

$$\overline{Y}^r_h \leq OM^r_{AV}, \qquad \forall r \in R^c, c \in IC, h \in \{0, 1, \ldots, \mathscr{H}^r\}. \tag{5.79}$$

$$OM^r_{AV} \leq \max_{r\in R^c, c\in IC}\left(\frac{-V^r_B+\ln(2n^r-1)}{\beta^c_0} - VOT^c_{AV} st^r\right), \qquad \forall r\in R^c, c\in IC \tag{5.80}$$

$$LN^r_{AV} - LN^r_B = V^r_{AV} - V^r_B, \qquad \forall r\in R^c, c\in IC \tag{5.81}$$

$$LN^r_{AV} \leq \frac{1}{u^k}P^r_{AV} + \ln u^k - 1, \qquad \forall r\in R^c, c\in IC, k\in K \tag{5.82}$$

$$LN^r_{AV} \geq \sum_{k=1}^{\mathscr{K}} \theta^k_r \ln u^k, \qquad \forall r\in R^c, c\in IC \tag{5.83}$$

$$P^r_{AV} = \sum_{k=1}^{\mathscr{K}} \theta^k_r u^k, \qquad \forall r\in R^c, c\in IC \tag{5.84}$$

$$\sum_{k=1}^{\mathscr{K}} \theta^k_r = 1, \qquad \forall r\in R^c, c\in IC \tag{5.85}$$

$$\sum_{k=1}^{\mathscr{K}-1} \lambda^k_r = 1, \qquad \forall r\in R^c, c\in IC \tag{5.86}$$

$$\theta^1_r \leq \lambda^1_r, \qquad \forall r\in R^c, c\in IC \tag{5.87}$$

$$\theta^k_r \leq \lambda^{k-1}_r + \lambda^k_r, \qquad \forall r\in R^c, c\in IC, k\in\{2,\ldots,\mathscr{K}-1\} \tag{5.88}$$

$$\theta^{\mathscr{K}}_r \leq \lambda^{\mathscr{K}-1}_r, \qquad \forall r\in R^c, c\in IC \tag{5.89}$$

$$LN^r_B \leq \frac{1}{u^k}(1-P^r_{AV}) + \ln u^k - 1, \qquad \forall r\in R^c, c\in IC, k\in K \tag{5.90}$$

$$LN^r_B \geq \sum_{k=1}^{\mathscr{K}} \overline{\theta}^k_r \ln u^k, \qquad \forall r\in R^c, c\in IC \tag{5.91}$$

$$1 - P^r_{AV} = \sum_{k=1}^{\mathscr{K}} \overline{\theta}^k_r u^k, \qquad \forall r\in R^c, c\in IC \tag{5.92}$$

$$\sum_{k=1}^{\mathscr{K}} \overline{\theta}^k_r = 1, \qquad \forall r\in R^c, c\in IC \tag{5.93}$$

$$\sum_{k=1}^{\mathscr{K}-1} \overline{\lambda}^k_r = 1, \qquad \forall r\in R^c, c\in IC \tag{5.94}$$

$$\overline{\theta}^1_r \leq \overline{\lambda}^1_r, \qquad \forall r\in R^c, c\in IC \tag{5.95}$$

$$\overline{\theta}^k_r \leq \overline{\lambda}^{k-1}_r + \overline{\lambda}^k_r, \qquad \forall r\in R^c, c\in IC, k\in\{2,\ldots,\mathscr{K}-1\} \tag{5.96}$$

$$\overline{\theta}_r^{\mathscr{K}} \leq \overline{\lambda}_r^{\mathscr{K}-1}, \qquad \forall r \in R^c, c \in IC \tag{5.97}$$

$$\sum_{r \in R^c, c \in IC} D_{AV}^r = \sum_{h=0}^{\mathscr{H}} 2^h \overline{D}_h \tag{5.98}$$

$$Y_h \leq \alpha, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{5.99}$$

$$Y_h \leq \overline{D}_h, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{5.100}$$

$$Y_h \geq \alpha + \overline{D}_h - 1, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{5.101}$$

$$\sum_{h=0}^{\mathscr{H}^r} 2^h Y_h = \sum_{r \in R^c, c \in IC} S^r \tag{5.102}$$

$$0 \leq \alpha \leq 1 \tag{5.103}$$

$$p^0 \geq 0 \tag{5.104}$$

$$p \geq 0 \tag{5.105}$$

$$p^c \geq 0, \qquad \forall c \in IC \tag{5.106}$$

$$\overline{D}_h \in \{0,1\}, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{5.107}$$

$$Y_h \geq 0, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}\} \tag{5.108}$$

$$\overline{S}_h^r \in \{0,1\}, \qquad \forall h \in \{0,1,\ldots,\mathscr{H}^r\} \tag{5.109}$$

$$\overline{Y}_h^r \geq 0, \qquad \forall r \in R^c, c \in IC, h \in \{0,1,\ldots,\mathscr{H}^r\} \tag{5.110}$$

$$V_{AV}^r \geq 0, \qquad \forall r \in R^c, c \in IC \tag{5.111}$$

$$T_{AV}^r \in \mathbb{N}^0, \qquad \forall r \in R^c, c \in IC \tag{5.112}$$

$$P_{AV}^r \geq 0, \qquad \forall r \in R^c, c \in IC \tag{5.113}$$

$$D_{AV}^r \in \mathbb{N}^0, \qquad \forall r \in R^c, c \in IC \tag{5.114}$$

$$O \in \mathbb{N}^0 \tag{5.115}$$

$$O_i \in \mathbb{N}^0, \qquad \forall i \in N_P \tag{5.116}$$

$$S^r \in \mathbb{N}^0, \qquad \forall r \in R^c, c \in IC \tag{5.117}$$

$$E_t^r \in \mathbb{N}^0, \qquad \forall r \in R^c, c \in IC, t \in T \tag{5.118}$$

$$PF_{i_{t_1} j_{t_2}}^r \in \mathbb{N}^0, \qquad \forall r \in R^c, c \in IC, (i_{t_1}, j_{t_2}) \in G \tag{5.119}$$

$$F_{i_{t_1} j_{t_2}} \geq 0, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{5.120}$$

$$W_{i_t} \in \mathbb{N}^0, \qquad \forall i \in N_P, t \in T \tag{5.121}$$

$$Z_t^r \in \{0,1\}, \qquad \forall r \in R^c, c \in IC, t \in T, a^r + st^r \leq b^r \tag{5.122}$$

$$X_{i_{t_1} j_{t_2}} \in \{0,1\}, \qquad \forall (i_{t_1}, j_{t_2}) \in G \tag{5.123}$$

$$A_t^r \in \{0,1\}, \qquad \forall r \in R^c, c \in IC, t \in T, a^r + st^r \leq t \leq b^r \tag{5.124}$$

$$LN_{AV}^r \geq 0, \qquad \forall r \in R^c, c \in IC \tag{5.125}$$

$$LN_B^r \geq 0, \qquad \forall r \in R^c, c \in IC \tag{5.126}$$

$$\theta_r^k \geq 0, \qquad \forall r \in R^c, c \in IC, k \in K \tag{5.127}$$

$$\lambda_r^k \in \{0,1\}, \qquad \forall r \in R^c, c \in IC, k \in K \tag{5.128}$$

$$\overline{\theta}_r^k \geq 0, \qquad \forall r \in R^c, c \in IC, k \in K \tag{5.129}$$

$$\overline{\lambda}_r^k \in \{0,1\}, \qquad \forall r \in R^c, c \in IC, k \in K \tag{5.130}$$

## 5.B Sensitivity analysis of parameters used in PSO

Parameter tuning is crucial for the performance of metaheuristic algorithms, as the selected settings can significantly influence the outcome. In this study, we conduct a sensitivity analysis on the key parameters of the PSO-based algorithm. These parameters include population size, the cognitive and social coefficients ($c_{1,\max}$ and $c_{2,\max}$), and the inertia coefficient ($w$). The scenarios with varying parameter settings are presented in Table 5.B.1. For each scenario, we conducted five experiments and reported the mean, minimum, and maximum values of the objective functions. These results are shown in Table 5.B.2, where only the objective function values are displayed as they serve as an indicator of model performance.

As shown in Table 5.B.2, varying the population size from 12 (base scenario) to 8 (Scenario 1) or 16 (Scenario 2) results in very similar overall performance, with Scenario 1 and 2 showing slightly worse performance in most cases. A population size of 8 reduces diversity within the population, while a population size of 16 increases the computational time per generation, leading to fewer iterations within the same timeframe. Larger cognitive and social coefficients ($c_{1,\max}$ and $c_{2,\max}$) promote more aggressive exploration and exploitation, whereas smaller values are more conservative and help prevent overshooting optimal regions. The results indicate that the performance in the base scenario and in Scenarios 3 and 4 is quite similar, making it difficult to draw a definitive conclusion about the impact of these coefficients on the outcome.

In Scenarios 5 and 6, we varied the inertia coefficient ($w$). A higher inertia encourages particles to retain a greater portion of their previous velocity, leading to more exploratory behaviour, with particles more likely to continue moving in their current direction. Conversely, lower inertia makes particles more responsive to the cognitive and social coefficients, thereby promoting quicker convergence and increased exploitation. The results from the base scenario, Scenario 5, and Scenario 6 are very similar. In our problem setting, these parameter variations do not significantly affect the model's performance, indicating that the algorithm is able to maintain stable performance across a range of different parameter settings.

*Table 5.B.1: Parameter setting for sensitivity analysis.*

| Scenario | Population size | $c_{1,\max}$,$c_{2,\max}$ | $w$ |
|---|---|---|---|
| Base scenario | 12 | 0.8 | 0.2 |
| Scenario 1 | 8 | 0.8 | 0.2 |
| Scenario 2 | 16 | 0.8 | 0.2 |
| Scenario 3 | 12 | 1 | 0.2 |
| Scenario 4 | 12 | 0.6 | 0.2 |
| Scenario 5 | 12 | 0.8 | 0.1 |
| Scenario 6 | 12 | 0.8 | 0.3 |

*Table 5.B.2: Objective function values across different scenarios under three pricing strategies.*

| | Pricing strategy 1 | | | Pricing strategy 2 | | | Pricing strategy 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| Base scenario | 7669.52 | 7553.52 | 7768.87 | 7161.29 | 7084.55 | 7207.06 | 8962.05 | 8778.77 | 9105.75 |
| Scenario 1 | 7547.43 | 7180.54 | 7760.51 | 7045.8 | 6931.4 | 7239.63 | 8822.34 | 8660 | 8920.54 |
| Scenario 2 | 7589.95 | 7513 | 7683.97 | 7090.33 | 6976.68 | 7241.6 | 8822.48 | 8737.16 | 8958.9 |
| Scenario 3 | 7597.93 | 7498.34 | 7790.28 | 6999.95 | 6398.17 | 7259.57 | 8814.33 | 8643.61 | 8963.8 |
| Scenario 4 | 7685.02 | 7590.01 | 7764.06 | 7158.50 | 6969.45 | 7254.77 | 8814.51 | 8659.01 | 9007.63 |
| Scenario 5 | 7654.72 | 7436.56 | 7797.37 | 6973.07 | 6423.31 | 7152.19 | 8910.99 | 8842.57 | 8974.48 |
| Scenario 6 | 7636.14 | 7549.43 | 7705.51 | 7182.61 | 7030.08 | 7289.55 | 8825.98 | 8733.11 | 8983.95 |

# Chapter 6

# Conclusions and future research

In this chapter, we first present the main conclusions in Section 6.1, summarizing the research questions, key findings, and overall conclusions. Section 6.2 offers recommendations for future research.

## 6.1 Conclusions

This thesis addresses the challenges of high-level planning for a ride-hailing service provider during the gradual transition from traditional transportation to a fully intelligent transportation system (ITS). To make optimal planning decisions, it is essential to understand and analyse system operations throughout a typical day. To this end, we have modelled the decision-making process at both the planning and operational levels. Given that urban demand patterns and transportation infrastructures are dynamic and constantly evolving, all decisions must be flexible and responsive to the current environment to maintain optimal profitability. This thesis provides mathematical models and designs various solution algorithms to tackle the challenges encountered during different stages of the transition period, thereby making the planning decisions more realistic. The research questions outlined in Chapter 1 have been systematically addressed throughout this thesis. Below, we summarise the answers to each question, along with the key findings, conclusions, and insights from case studies.

*Research question 1: How should ride-hailing service providers optimally size and manage mixed fleets of SAVs and conventional vehicles/taxis in response to the gradual expansion of AV-only zones in urban areas, considering their impact on traffic congestion?*

To answer this question, we developed a mixed-integer linear programming model in Chapter 2 to determine the optimal fleet size and type across different service sce-

narios and to evaluate how an AVs-only zone affects the service performance and the planning decisions. Traffic congestion is incorporated into the model through flow-dependent travel times.

Our findings indicate that automated taxis (ATs) generally yield higher profits compared to conventional taxis (CTs). If the passenger preference for a vehicle type is not considered, the operating company should increase the deployment of ATs in response to the introduction of an AVs-only zone. When the service provider considers the user's preference towards the vehicle types, less profit will be gained.

Initially, the establishment of an AVs-only zone may increase detours and relocation distances for CTs. However, a well-designed expansion strategy for the AVs-only zone can help mitigate these negative impacts. As the coverage of the AVs-only zone expands, traffic congestion is likely to decrease, enabling the company to achieve higher profit by deploying more ATs.

*Research question 2: How can we model the interactions between different routing behaviours—specifically, privately-owned HVs following the user equilibrium (UE) and centrally dispatched vehicles/taxis following the system optimum (SO)? How do these interactions influence the optimal sizing and management of the fleets?*

To answer this question, we developed a bi-level framework in Chapter 3. The lower level of this framework presents an approximated dynamic mixed equilibrium model designed to capture the interactions between vehicles with different routing behaviours—privately-owned HVs operating under UE and centrally dispatched vehicles/taxis (CTs and ATs) following an SO routing strategy. At the upper level, the objective is to determine the fleet size of CTs and ATs that maximises the profits of the company while meeting travel demand. A parallel genetic algorithm is developed to solve the proposed bi-level framework, embedded with a tailored iterative algorithm to solve the lower-level problem.

Computational experiments conducted using the city of Delft as a case study demonstrate the effectiveness of this approach in determining near-optimal fleet sizes of CTs and ATs under different scenarios with varying departure times and user preferences for CTs and ATs. However, applying this solution method to large urban networks can lead to a long computational time due to the NP-hardness of the model and the iterative nature of the framework.

Several key findings are drawn from the experiments. First, the minimum fleet size required to meet demand is not necessarily the most profitable fleet size for a ride-hailing company. Among all expenses associated with CTs, driver salaries represent a substantial cost, which heavily influences fleet size decisions. As a result, the minimum feasible fleet size for CTs often aligns with the optimal choice in the scenarios tested. In addition, the location and distribution of parking depots also plays a crucial

role in influencing the fleet sizes, with depots located in high-demand areas helping to minimize relocation costs. Second, AVs-only zones can enhance transportation efficiency by reducing congestion, though this benefit is less noticeable in the early stages of implementation. To maximise the benefits of AVs-only zones, governments should encourage the adoption of AVs.

*Research question 3: How can existing models be adapted to incorporate endogenous demand to plan and operate an SAV service?*

To answer this question, we present a non-convex, non-linear mathematical programming model in Chapter 4. In this model, we model travellers' mode choice behaviour between SAVs and bicycles using an endogenous binary logit model, under the assumption that private cars are banned in urban areas. The binary logit model is incorporated into a mixed-integer programming model which aims at optimising fleet sizing and management decisions for an SAV service. This model takes into account traffic congestion, the non-linear demand of users across different income classes, and different accept/reject mechanisms which influence travellers' willingness to use the SAV service.

To address the computational challenges posed by the model's non-linearities, we reformulated the problem and applied outer-inner approximation methods along with a breakpoint generation algorithm to obtain an approximated linear version of the original model. This allows the model to be solved using advanced solvers like Gurobi.

We conducted a quasi-real case study in Delft, the Netherlands, accompanied by a sensitivity analysis, to evaluate the model's performance and provide practical insights for SAV service providers in future scenarios. The results reveal that demand for SAVs, supply strategies, and network performance, particularly traffic congestion, are deeply interconnected. This highlights the importance of considering these interactions when managing SAV fleets.

In terms of fleet sizing strategy, we conclude that factors such as population distribution, land use patterns, and residents' travel behaviour significantly influence the initial distribution of the SAV fleet. Additionally, the location and distribution of parking depots are crucial in determining fleet sizes, with depots in high-demand areas helping to minimise relocation costs.

The study also finds that SAV services are more appealing to travellers with a higher value of travel time (VOTT), who are more sensitive to trip length/duration variations. For long trips, these travellers consistently prefer SAVs, while those with lower VOTT favour SAVs only when prices are low. For shorter trips, bicycles are generally preferred unless SAV prices are significantly reduced.

*Research question 4: What are the optimal pricing strategies for SAV services, considering the interplay between demand and supply variables, congestion effects, and the heterogeneous income levels of travellers?*

The thesis evaluates three pricing strategies in Chapter 5: base fare plus distance-based fare, distance-based fare only, and income class-based fare. To analyse optimal pricing strategies while considering demand-supply interplay, congestion effects and the heterogeneous income levels of travellers, we develop a mixed-integer nonlinear programming model. To solve the problem, we propose three solution methods. The first method involves linearising the model and solving it using a commercial solver. The second approach combines Particle Swarm Optimisation with an iterative framework to address a reformulated mixed-integer linear programming model. The third method integrates Bayesian optimisation with the reformulated mixed-integer linear programming model to handle complex objective function evaluations. These methods are applied to both a small-scale case study and a larger case study in Delft, the Netherlands.

The results show that the three pricing strategies have distinctly different impacts on demand and profitability. The income class-based fare generates the highest profit for the SAV service provider by attracting greater overall demand. This strategy charges higher rates to high-income travellers, leveraging their lower price sensitivity to maximise revenue. While it leads to increased congestion, it also encourages usage among low- and middle-income users, promoting social equity while remaining the most profitable.

In contrast, the base fare plus distance-based fare proves to be the most environmentally sustainable, as it discourages SAV use for short trips (less than 5 km) and promotes active transport. The distance-based fare results in lower congestion and fewer SAVs on the road but yields the lowest profit. Traveller behaviour varies significantly by socio-demographic factors, with demand for SAV services increasing with income, particularly for longer trips. High-income individuals are more likely to opt for SAVs.

## 6.2 Future research

The findings and methodologies presented in this thesis open several directions for future research.

### 6.2.1 Methodological outlook

Firstly, future work could focus on incorporating stochastic elements such as demand uncertainty, departure time variability, and travel time fluctuations into fleet sizing models. These factors play a critical role in real-world operations, where demand and travel conditions are inherently unpredictable. The absence of these stochastic elements in current models could lead to overly deterministic results, which may not fully capture the complexities of urban transportation systems. By integrating stochastic elements, future models could provide more robust solutions that account for a wider range of possible scenarios. This would improve the decision-making process by offering fleet sizing strategies that are resilient to variability and uncertainty.

Secondly, there is a methodological need to optimise the design of AV-only zones across multiple time periods, taking into account travellers' mode choices, routing behaviours, and congestion effects. This requires advanced methodologies capable of modelling time-dependent lane and link transitions, strategically allocating AV parking depots, and capturing the changes in travellers' behaviour over time.

Thirdly, investigating the competitive dynamics among multiple SAV service providers through advanced modelling techniques could provide a deeper understanding of market interactions. For instance, this could be approached as a multi-agent reinforcement learning problem, where service providers dynamically adjust pricing strategies in response to competitors and evolving market conditions.

### 6.2.2 Practical outlook

From a practical perspective, future research should focus on optimising the interaction between SAVs and traditional public transportation systems such as buses, metro, and trains. Understanding these interactions will be crucial for creating an integrated urban transportation network. Specifically, practical research should explore how planning and operational decisions in one mode affect others, including potential conflicts like competition for passengers and the substitution effects of SAVs on public transit.

Another practical area is the design and implementation of AV-only zones. While methodological frameworks can guide the optimisation, future research must address real-world deployment challenges such as public acceptance, regulatory constraints, and infrastructure readiness.

Lastly, understanding the implications of dynamic pricing on passenger satisfaction, behaviour, and overall network efficiency in real-world scenarios is important. Field studies or pilot programs could validate the theoretical models and help refine pricing strategies based on observed outcomes.

# Bibliography

Al-Kanj, L., J. Nascimento, W. B. Powell (2020) Approximate dynamic programming for planning a ride-hailing system using autonomous fleets of electric vehicles, *European Journal of Operational Research*, 284(3), pp. 1088–1106.

Allahviranloo, M., J. Y. Chow (2019) A fractionally owned autonomous vehicle fleet sizing problem with time slot demand substitution effects, *Transportation Research Part C: Emerging Technologies*, 98, pp. 37–53.

Asadpour, A., I. Lobel, G. van Ryzin (2023) Minimum earnings regulation and the stability of marketplaces, *Manufacturing & Service Operations Management*, 25(1), pp. 254–265.

Ashkrof, P., G. H. de Almeida Correia, O. Cats, B. Van Arem (2022a) Ride acceptance behaviour of ride-sourcing drivers, *Transportation Research Part C: Emerging Technologies*, 142, p. 103783.

Ashkrof, P., G. Homem Correia, O. Cats, B. Van Arem (2022b) Ride acceptance behaviour of ride-sourcing drivers, *Transportation Research Part C: Emerging Technologies*, 142, p. 103783.

Ashkrof, P., G. Homem de Almeida Correia, O. Cats, B. van Arem (2019) Impact of automated vehicles on travel mode preference for different trip purposes and distances, *Transportation Research Record*, 2673(5), pp. 607–616.

Atasoy, B., M. Salani, M. Bierlaire (2014) An integrated airline scheduling, fleeting, and pricing model for a monopolized market, *Computer-Aided Civil and Infrastructure Engineering*, 29(2), pp. 76–90.

Azadeh, S. S., J. van der Zee, M. Wagenvoort (2022) Choice-driven service network design for an integrated fixed line and demand responsive mobility system, *Transportation Research Part A: Policy and Practice*, 166, pp. 557–574.

Bagloee, S. A., M. Sarvi, M. Patriksson, A. Rajabifard (2017) A mixed user-equilibrium and system-optimal traffic flow for connected vehicles stated as a complementarity problem, *Computer-Aided Civil and Infrastructure Engineering*, 32(7), pp. 562–580.

Bai, J., C. S. Tang (2022) Can two competing on-demand service platforms be profitable?, *International Journal of Production Economics*, 250, p. 108672.

Balac, M., S. Hörl, K. W. Axhausen (2020) Fleet sizing for pooled (automated) vehicle fleets, *Transportation Research Record*, 2674(9), pp. 168–176.

Beirigo, B. A., F. Schulte, R. R. Negenborn (2022) A learning-based optimization approach for autonomous ridesharing platforms with service-level contracts and on-demand hiring of idle vehicles, *Transportation Science*, 56(3), pp. 677–703, publisher: INFORMS.

Ben-Akiva, M. E., S. R. Lerman, S. R. Lerman, et al. (1985) *Discrete choice analysis: theory and application to travel demand*, vol. 9, MIT press.

Bergstra, J., R. Bardenet, Y. Bengio, B. Kégl (2011) Algorithms for hyper-parameter optimization, *Advances in Neural Information Processing Systems*, 24.

BicycleDutch (2018) Dutch cycling figures, URL https://bicycledutch.wordpress.com/2018/01/02/dutch-cycling-figures.

Bösch, P. M., F. Becker, H. Becker, K. W. Axhausen (2018) Cost-based analysis of autonomous mobility services, *Transport Policy*, 64, pp. 76–91.

Brandão, J. (2009) A deterministic tabu search algorithm for the fleet size and mix vehicle routing problem, *European Journal of Operational Research*, 195(3), pp. 716–728.

Cachon, G. P., K. M. Daniels, R. Lobel (2017) The role of surge pricing on a service platform with self-scheduling capacity, *Manufacturing & Service Operations Management*, 19(3), pp. 368–384.

Cadarso, L., V. Vaze, C. Barnhart, Á. Marín (2017) Integrated airline scheduling: Considering competition effects and the entry of the high speed rail, *Transportation Science*, 51(1), pp. 132–154.

Cai, Y., J. Chen, D. Lei, J. Yu, et al. (2022) The integration of multimodal networks: The generalized modal split and collaborative optimization of transportation hubs, *Journal of Advanced Transportation*, 2022.

Camacho-Vallejo, J.-F., L. López-Vera, A. E. Smith, J.-L. González-Velarde (2021) A tabu search algorithm to solve a green logistics bi-objective bi-level problem, *Annals of Operations Research*, pp. 1–27.

Chen, A., S. Pravinvongvuth, X. Xu, S. Ryu, P. Chootinan (2012) Examining the scaling effect and overlapping problem in logit-based stochastic user equilibrium models, *Transportation Research Part A: Policy and Practice*, 46(8), pp. 1343–1358.

Chen, R., M. W. Levin (2019) Dynamic user equilibrium of mobility-on-demand system with linear programming rebalancing strategy, *Transportation Research Record*, 2673(1), pp. 447–459.

Chen, X. M., H. Zheng, J. Ke, H. Yang (2020) Dynamic optimization strategies for on-demand ride services platform: Surge pricing, commission rate, and incentives, *Transportation Research Part B: Methodological*, 138, pp. 23–45.

Chen, Z., F. He, Y. Yin, Y. Du (2017) Optimal design of autonomous vehicle zones in transportation networks, *Transportation Research Part B: Methodological*, 99, pp. 44–61.

Chen, Z., F. He, L. Zhang, Y. Yin (2016) Optimal deployment of autonomous vehicle lanes with endogenous market penetration, *Transportation Research Part C: Emerging Technologies*, 72, pp. 143–156.

Chondrogiannis, T., P. Bouros, J. Gamper, U. Leser, D. B. Blumenthal (2020) Finding k-shortest paths with limited overlap, *The VLDB Journal*, 29(5), pp. 1023–1047.

Chung, J.-H., K. Y. Hwang, Y. K. Bae (2012) The loss of road capacity and self-compliance: Lessons from the Cheonggyecheon stream restoration, *Transport Policy*, 21, pp. 165–178.

Conceição, L., G. Correia, J. P. Tavares (2021) Automated vehicles (AV) dedicated networks and their effects on the traveling of conventional vehicle drivers, *Transportation Research Procedia*, 52, pp. 653–660.

Correia, G. H., E. Looff, S. Van Cranenburgh, M. Snelder, B. Van Arem (2019) On the impact of vehicle automation on the value of travel time while performing work and leisure activities in a car: Theoretical insights and results from a stated preference survey, *Transportation Research Part A: Policy and Practice*, 119, pp. 359–382.

Correia, G. H., B. Van Arem (2016) Solving the user optimum privately owned automated vehicles assignment problem (UO-POAVAP): A model to explore the impacts of self-driving vehicles on urban mobility, *Transportation Research Part B: Methodological*, 87, pp. 64–88.

Dafermos, S. C., F. T. Sparrow (1969) The traffic assignment problem for a general network, *Journal of Research of the National Bureau of Standards B*, 73(2), pp. 91–118.

Delft High Performance Computing Centre (DHPC) (2024) DelftBlue Supercomputer (Phase 2), https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2.

Dong, X., J. Y. Chow, S. T. Waller, D. Rey (2022) A chance-constrained dial-a-ride problem with utility-maximising demand and multiple pricing structures, *Transportation Research Part E: Logistics and Transportation Review*, 158, p. 102601.

Eklund, S. E. (2004) A massively parallel architecture for distributed genetic algorithms, *Parallel Computing*, 30(5-6), pp. 647–676.

Fagnant, D. J., K. M. Kockelman (2014) The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios, *Transportation Research Part C: Emerging Technologies*, 40, pp. 1–13.

Fagnant, D. J., K. M. Kockelman (2018) Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas, *Transportation*, 45(1), pp. 143–158.

Fan, Q., J. T. Van Essen, G. H. Correia (2022) Heterogeneous fleet sizing for on-demand transport in mixed automated and non-automated urban areas, *Transportation Research Procedia*, 62, pp. 163–170.

Fan, Q., J. T. van Essen, G. H. Correia (2023) Optimising fleet sizing and management of shared automated vehicle (SAV) services: A mixed-integer programming approach integrating endogenous demand, congestion effects, and accept/reject mechanism impacts, *Transportation Research Part C: Emerging Technologies*, 157, p. 104398.

Farahani, R. Z., E. Miandoabchi, W. Y. Szeto, H. Rashidi (2013) A review of urban transportation network design problems, *European Journal of Operational Research*, 229(2), pp. 281–302.

Ge, Q., K. Han, X. Liu (2021) Matching and routing for shared autonomous vehicles in congestible network, *Transportation Research Part E: Logistics and Transportation Review*, 156, p. 102513.

Guo, H., Y. Chen, Y. Liu (2022) Shared autonomous vehicle management considering competition with human-driven private vehicles, *Transportation Research Part C: Emerging Technologies*, 136, p. 103547.

Guo, Q., X. J. Ban, H. A. Aziz (2021a) Mixed traffic flow of human driven vehicles and automated vehicles on dynamic transportation networks, *Transportation Research Part C: Emerging Technologies*, 128, p. 103159.

Guo, Z., M. Hao, B. Yu, B. Yao (2021b) Robust minimum fleet problem for autonomous and human-driven vehicles in on-demand ride services considering mixed operation zones, *Transportation Research Part C: Emerging Technologies*, 132, p. 103390.

Gurumurthy, K. M., F. de Souza, A. Enam, J. Auld (2020) Integrating supply and demand perspectives for a large-scale simulation of shared autonomous vehicles, *Transportation Research Record*, 2674(7), pp. 181–192.

Ha, T., S. Kim, D. Seo, S. Lee (2020) Effects of explanation types and perceived risk on trust in autonomous vehicles, *Transportation Research Part F: Traffic Psychology and Behaviour*, 73, pp. 271–280.

Hiermann, G., J. Puchinger, S. Ropke, R. F. Hartl (2016) The electric fleet size and mix vehicle routing problem with time windows and recharging stations, *European Journal of Operational Research*, 252(3), pp. 995–1018.

Hoang, N. H., M. Panda, H. L. Vu, D. Ngoduy, H. K. Lo (2023) A new framework for mixed-user dynamic traffic assignment considering delay and accessibility to information, *Transportation Research Part C: Emerging Technologies*, 146, p. 103977.

Hörl, S., F. Becker, K. W. Axhausen (2021) Simulation of price, customer behaviour and system impact for a cost-covering automated taxi system in Zurich, *Transportation Research Part C: Emerging Technologies*, 123, p. 102974.

Huang, K., K. An, J. Rich, W. Ma (2020) Vehicle relocation in one-way station-based electric carsharing systems: A comparative study of operator-based and user-based methods, *Transportation Research Part E: Logistics and Transportation Review*, 142, p. 102081.

Huang, K., G. H. de Almeida Correia, K. An (2018) Solving the station-based one-way carsharing network planning problem with relocations and non-linear demand, *Transportation Research Part C: Emerging Technologies*, 90, pp. 1–17.

Huang, Y., K. M. Kockelman (2020) Electric vehicle charging station locations: Elastic demand, station congestion, and network equilibrium, *Transportation Research Part D: Transport and Environment*, 78, p. 102179.

Hulse, L. M. (2023) Pedestrians' perceived vulnerability and observed behaviours relating to crossing and passing interactions with autonomous vehicles, *Transportation Research Part F: Traffic Psychology and Behaviour*, 93, pp. 34–54.

Hyland, M. F., H. S. Mahmassani (2017) Taxonomy of shared autonomous vehicle fleet management problems to inform future transportation mobility, *Transportation Research Record*, 2653(1), pp. 26–34.

Joksimovic, D., M. C. Bliemer, P. H. Bovy (2005) Optimal toll design problem in dynamic traffic networks with joint route and departure time choice, *Transportation Research Record*, 1923(1), pp. 61–72.

Jorge, D., G. Molnar, G. H. de Almeida Correia (2015) Trip pricing of one-way station-based carsharing networks with zone and time of day price variations, *Transportation Research Part B: Methodological*, 81, pp. 461–482.

Kashmiri, F. A., H. K. Lo (2022) Routing of autonomous vehicles for system optimal flows and average travel time equilibrium over time, *Transportation Research Part C: Emerging Technologies*, 143, p. 103818.

Katoch, S., S. S. Chauhan, V. Kumar (2021) A review on genetic algorithm: past, present, and future, *Multimedia Tools and Applications*, 80(5), pp. 8091–8126.

Kaufman, D. E., J. Nonis, R. L. Smith (1998) A mixed integer linear programming model for dynamic route guidance, *Transportation Research Part B: Methodological*, 32(6), pp. 431–440.

Ke, Z., S. Qian (2023) Leveraging ride-hailing services for social good: Fleet optimal routing and system optimal pricing, *Transportation Research Part C: Emerging Technologies*, 155, p. 104284.

Koç, Ç., T. Bektaş, O. Jabali, G. Laporte (2016) The fleet size and mix location-routing problem with time windows: Formulations and a heuristic algorithm, *European Journal of Operational Research*, 248(1), pp. 33–51.

Kolarova, V., F. Steck, F. J. Bahamonde-Birke (2019) Assessing the effect of autonomous driving on value of travel time savings: A comparison between current and future preferences, *Transportation Research Part A: Policy and Practice*, 129, pp. 155–169.

Kouwenhoven, M., G. C. de Jong, P. Koster, V. A. Van den Berg, E. T. Verhoef, J. Bates, P. M. Warffemius (2014) New values of time and reliability in passenger transport in the Netherlands, *Research in Transportation Economics*, 47, pp. 37–49.

Laporte, G. (2009) Fifty years of vehicle routing, *Transportation Science*, 43(4), pp. 408–416.

Li, Q., F. Liao (2020) Incorporating vehicle self-relocations and traveler activity chains in a bi-level model of optimal deployment of shared autonomous vehicles, *Transportation Research Part B: Methodological*, 140, pp. 151–175.

Li, R., X. Liu, Y. M. Nie (2018) Managing partially automated network traffic flow: Efficiency vs. stability, *Transportation Research Part B: Methodological*, 114, pp. 300–324.

Li, S., H. Yang, K. Poolla, P. Varaiya (2021) Spatial pricing in ride-sourcing markets under a congestion charge, *Transportation Research Part B: Methodological*, 152, pp. 18–45.

Liang, Q., X.-a. Li, Z. Chen, T. Pan, R. Zhong (2023) Day-to-day traffic control for networks mixed with regular human-piloted and connected autonomous vehicles, *Transportation Research Part B: Methodological*, 178, p. 102847.

Liang, X., G. H. Correia, K. An, B. Van Arem (2020) Automated taxis' dial-a-ride problem with ride-sharing considering congestion-based dynamic travel times, *Transportation Research Part C: Emerging Technologies*, 112, pp. 260–281.

Liang, X., G. H. Correia, B. Van Arem (2017) An optimization model for vehicle routing of automated taxi trips with dynamic travel times, *Transportation Research Procedia*, 27, pp. 736–743.

Liang, X., G. Homem Correia, B. Van Arem (2018) Applying a model for trip assignment and dynamic routing of automated taxis with congestion: system performance in the city of Delft, the Netherlands, *Transportation Research Record*, 2672(8), pp. 588–598.

Lin, X., Y.-W. Zhou (2019) Pricing policy selection for a platform providing vertically differentiated services with self-scheduling capacity, *Journal of the Operational Research Society*, 70(7), pp. 1203–1218.

Liu, H., C. Jin, B. Yang, A. Zhou (2017) Finding top-k shortest paths with diversity, *IEEE Transactions on Knowledge and Data Engineering*, 30(3), pp. 488–502.

Liu, H., D. Z. Wang (2015) Global optimization method for network design problem with stochastic user equilibrium, *Transportation Research Part B: Methodological*, 72, pp. 20–39.

Liu, J., P. Mirchandani, X. Zhou (2020) Integrated vehicle assignment and routing for system-optimal shared mobility planning with endogenous road congestion, *Transportation Research Part C: Emerging Technologies*, 117, p. 102675.

Liu, S., W. Huang, H. Ma (2009) An effective genetic algorithm for the fleet size and mix vehicle routing problems, *Transportation Research Part E: Logistics and Transportation Review*, 45(3), pp. 434–445.

Liu, Y., P. Bansal, R. Daziano, S. Samaranayake (2019) A framework to integrate mode choice in the design of mobility-on-demand systems, *Transportation Research Part C: Emerging Technologies*, 105, pp. 648–665.

Liu, Z., Z. Song (2019) Strategic planning of dedicated autonomous vehicle lanes and autonomous vehicle/toll lanes in transportation networks, *Transportation Research Part C: Emerging Technologies*, 106, pp. 381–403.

Lou, Y., Y. Yin, J. A. Laval (2011) Optimal dynamic pricing strategies for high-occupancy/toll lanes, *Transportation Research Part C: Emerging Technologies*, 19(1), pp. 64–74.

Lu, R., G. H. d. A. Correia, X. Zhao, X. Liang, Y. Lv (2021) Performance of one-way carsharing systems under combined strategy of pricing and relocations, *Transportmetrica B: Transport Dynamics*, 9(1), pp. 134–152.

Madadi, B., R. Van Nes, M. Snelder, B. Van Arem (2020) A bi-level model to optimize road networks for a mixture of manual and automated driving: An evolutionary local search algorithm, *Computer-Aided Civil and Infrastructure Engineering*, 35(1), pp. 80–96.

Madigan, R., S. Nordhoff, C. Fox, R. E. Amini, T. Louw, M. Wilbrink, A. Schieben, N. Merat (2019) Understanding interactions between automated road transport systems and other road users: A video analysis, *Transportation Research Part F: Traffic Psychology and Behaviour*, 66, pp. 196–213.

UITP (2017) Autonomous vehicles: a potential game changer for urban mobility. in: Policy brief. brussels: International Association of Public Transport (UITP).

Mansourianfar, M. H., Z. Gu, M. Saberi (2022) Distance-based time-dependent optimal ratio control scheme (TORCS) in congested mixed autonomy networks, *Transportation Research Part C: Emerging Technologies*, 141, p. 103760.

Mansourianfar, M. H., Z. Gu, S. T. Waller, M. Saberi (2021) Joint routing and pricing control in congested mixed autonomy networks, *Transportation Research Part C: Emerging Technologies*, 131, p. 103338.

Meskar, M., S. Aslani, M. Modarres (2023) Spatio-temporal pricing algorithm for ride-hailing platforms where drivers can decline ride requests, *Transportation Research Part C: Emerging Technologies*, 153, p. 104200.

Militão, A. M., A. Tirachini (2021) Optimal fleet size for a shared demand-responsive transport system with human-driven vs automated vehicles: A total cost minimization approach, *Transportation Research Part A: Policy and Practice*, 151, pp. 52–80.

Mo, D., X. M. Chen, J. Zhang (2022) Modeling and managing mixed on-demand ride services of human-driven vehicles and autonomous vehicles, *Transportation Research Part B: Methodological*, 157, pp. 80–119.

Müller, C., J. Gönsch, M. Soppert, C. Steinhardt (2023) Customer-centric dynamic pricing for free-floating vehicle sharing systems, *Transportation Science*, 57(6), pp. 1406–1432.

Nieuwenhuijsen, M. J., H. Khreis (2016) Car free cities: Pathway to healthy urban living, *Environment International*, 94, pp. 251–262.

Nourinejad, M., M. Ramezani (2020) Ride-sourcing modeling and pricing in non-equilibrium two-sided markets, *Transportation Research Part B: Methodological*, 132, pp. 340–357.

Oh, S., R. Seshadri, C. L. Azevedo, N. Kumar, K. Basak, M. Ben-Akiva (2020) Assessing the impacts of automated mobility-on-demand through agent-based simulation:

A study of Singapore, *Transportation Research Part A: Policy and Practice*, 138, pp. 367–388.

Olia, A., S. Razavi, B. Abdulhai, H. Abdelgawad (2018) Traffic capacity implications of automated vehicles mixed with regular vehicles, *Journal of Intelligent Transportation Systems*, 22(3), pp. 244–262.

On-Road Automated Driving (ORAD) committee (2021) *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International.

Özkan, E. (2020) Joint pricing and matching in ride-sharing systems, *European Journal of Operational Research*, 287(3), pp. 1149–1160.

Paneque, M. P., M. Bierlaire, B. Gendron, S. S. Azadeh (2021) Integrating advanced discrete choice models in mixed integer linear optimization, *Transportation Research Part B: Methodological*, 146, pp. 26–49.

Paneque, M. P., B. Gendron, S. S. Azadeh, M. Bierlaire (2022) A Lagrangian decomposition scheme for choice-based optimization, *Computers & Operations Research*, 148, p. 105985.

Pinto, H. K. R. F., M. F. Hyland, H. S. Mahmassani, I. O. Verbas (2020) Joint design of multimodal transit networks and shared autonomous mobility fleets, *Transportation Research Part C: Emerging Technologies*, 113, pp. 2–20.

Ren, X., J. Y. Chow (2022) A random-utility-consistent machine learning method to estimate agents' joint activity scheduling choice from a ubiquitous data set, *Transportation Research Part B: Methodological*, 166, pp. 396–418.

Renaud, J., F. F. Boctor (2002) A sweep-based algorithm for the fleet size and mix vehicle routing problem, *European Journal of Operational Research*, 140(3), pp. 618–628.

Santos, G. G. D., G. M. Correia (2021) A flow-based integer programming approach to design an interurban shared automated vehicle system and assess its financial viability, *Transportation Research Part C: Emerging Technologies*, 128, p. 103092.

Scherr, Y. O., B. A. N. Saavedra, M. Hewitt, D. C. Mattfeld (2019) Service network design with mixed autonomous fleets, *Transportation Research Part E: Logistics and Transportation Review*, 124, pp. 40–55.

Schnabel, W., D. Löhse (1997) Grundlagen der strasssen-verkehrstechnik und der verkehrsplanung, band 2, neue bearb.

Sheffi, Y. (1985) *Urban transportation networks*, vol. 6, Prentice-Hall, Englewood Cliffs, NJ.

Spieser, K., K. Treleaven, R. Zhang, E. Frazzoli, D. Morton, M. Pavone (2014) Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore, *Road Vehicle Automation*, pp. 229–245.

Steiner, K., S. Irnich (2020) Strategic planning for integrated mobility-on-demand and urban public bus networks, *Transportation Science*, 54(6), pp. 1616–1639.

Stoiber, T., I. Schubert, R. Hoerler, P. Burger (2019) Will consumers prefer shared and pooled-use autonomous vehicles? A stated choice experiment with Swiss households, *Transportation Research Part D: Transport and Environment*, 71, pp. 265–282.

Swersky, K., J. Snoek, R. P. Adams (2013) Multi-task Bayesian optimization, *Advances in Neural Information Processing Systems*, 26.

Taylor, T. A. (2018) On-demand service platforms, *Manufacturing & Service Operations Management*, 20(4), pp. 704–720.

Tennøy, A. (2010) Why we fail to reduce urban road traffic volumes: Does it matter how planners frame the problem?, *Transport Policy*, 17(4), pp. 216–223.

Tian, J., H. Jia, G. Wang, Y. Lin, R. Wu, A. Lv (2022) A long-term shared autonomous vehicle system design problem considering relocation and pricing, *Journal of Advanced Transportation*, 2022.

Tong, Y., L. Wang, Z. Zhou, L. Chen, B. Du, J. Ye (2018) Dynamic pricing in spatial crowdsourcing: A matching-based approach, in: *Proceedings of the 2018 International Conference on Management of Data*, pp. 773–788.

Uber (2023) Uber, URL https://www.uber.com/.

Van Essen, J. T., G. H. Correia (2019) Exact formulation and comparison between the user optimum and system optimum solution for routing privately owned automated vehicles, *IEEE Transactions on Intelligent Transportation Systems*, 20(12), pp. 4567–4578.

Västberg, O. B., A. Karlström, D. Jonsson, M. Sundberg (2020) A dynamic discrete choice activity-based travel demand model, *Transportation Science*, 54(1), pp. 21–41.

Verbich, D., A. El-Geneidy (2017) Public transit fare structure and social vulnerability in Montreal, Canada, *Transportation Research Part A: Policy and Practice*, 96, pp. 43–53.

Vlakveld, W., S. van der Kint, M. P. Hagenzieker (2020) Cyclists' intentions to yield for automated cars at intersections when they have right of way: Results of an experiment using high-quality video animations, *Transportation Research Part F: Traffic Psychology and Behaviour*, 71, pp. 288–307.

Wang, D. Z., H. Liu, W. Szeto (2015) A novel discrete network design problem formulation and its global optimization solution algorithm, *Transportation Research Part E: Logistics and Transportation Review*, 79, pp. 213–230.

Wang, D. Z., H. K. Lo (2010) Global optimum of the linearized network design problem with equilibrium flows, *Transportation Research Part B: Methodological*, 44(4), pp. 482–492.

Wang, J., L. Zhang, Y. Huang, J. Zhao (2020a) Safety of autonomous vehicles, *Journal of Advanced Transportation*, 2020(1), p. 8867757.

Wang, S., G. H. d. A. Correia, H. X. Lin (2022a) Assessing the potential of the strategic formation of urban platoons for shared automated vehicle fleets, *Journal of Advanced Transportation*, 2022(1), p. 1005979.

Wang, S., G. H. de Almeida Correia, H. X. Lin (2022b) Modeling the competition between multiple automated mobility on-demand operators: An agent-based approach, *Physica A: Statistical Mechanics and its Applications*, 605, p. 128033.

Wang, S., B. Mo, J. Zhao (2020b) Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions, *Transportation Research Part C: Emerging Technologies*, 112, pp. 234–251.

Wang, X., F. He, H. Yang, H. O. Gao (2016) Pricing strategies for a taxi-hailing platform, *Transportation Research Part E: Logistics and Transportation Review*, 93, pp. 212–231.

Wang, Z., M. Qi, C. Cheng, C. Zhang (2019) A hybrid algorithm for large-scale service network design considering a heterogeneous fleet, *European Journal of Operational Research*, 276(2), pp. 483–494.

Wei, K., V. Vaze, A. Jacquillat (2022) Transit planning optimization under ride-hailing competition and traffic congestion, *Transportation Science*, 56(3), pp. 725–749.

Wu, T., M. Zhang, X. Tian, S. Wang, G. Hua (2020) Spatial differentiation and network externality in pricing mechanism of online car hailing platform, *International Journal of Production Economics*, 219, pp. 275–283.

Xu, M., Q. Meng (2020) Optimal deployment of charging stations considering path deviation and nonlinear elastic demand, *Transportation Research Part B: Methodological*, 135, pp. 120–142.

Xu, M., Q. Meng, Z. Liu (2018a) Electric vehicle fleet size and trip pricing for one-way carsharing services considering vehicle relocation and personnel assignment, *Transportation Research Part B: Methodological*, 111, pp. 60–82.

Xu, X., A. Chen, S. Jansuwan, C. Yang, S. Ryu (2018b) Transportation network redundancy: Complementary measures and computational methods, *Transportation Research Part B: Methodological*, 114, pp. 68–85.

Yang, K., S. I. Guler, M. Menendez (2016) Isolated intersection control for various levels of vehicle technology: Conventional, connected, and automated vehicles, *Transportation Research Part C: Emerging Technologies*, 72, pp. 109–129.

Yang, K., M. W. Tsao, X. Xu, M. Pavone (2020) Planning and operations of mixed fleets in mobility-on-demand systems, *arXiv preprint arXiv:2008.08131*.

Yang, S., J. Wu, H. Sun, Y. Qu, D. Z. W. Wang (2022) Integrated optimization of pricing and relocation in the competitive carsharing market: A multi-leader-follower game model, *Transportation Research Part C: Emerging Technologies*, 138, p. 103613.

Ye, J., Y. Jiang, J. Chen, Z. Liu, R. Guo (2021) Joint optimisation of transfer location and capacity for a capacitated multimodal transport network with elastic demand: a bi-level programming model and paradoxes, *Transportation Research Part E: Logistics and Transportation Review*, 156, p. 102540.

Yen, J. Y. (1970) An algorithm for finding shortest routes from all source nodes to a given destination in general networks, *Quarterly of Applied Mathematics*, 27(4), pp. 526–530.

Yi, Z., J. Smart (2021) A framework for integrated dispatching and charging management of an autonomous electric vehicle ride-hailing fleet, *Transportation Research Part D: Transport and Environment*, 95, p. 102822.

You, L., J. He, J. Zhao, J. Xie (2022) A federated mixed logit model for personal mobility service in autonomous transportation systems, *Systems*, 10(4), p. 117.

Zhang, F., J. Lu, X. Hu (2022) Integrated path controlling and subsidy scheme for mobility and environmental management in automated transportation networks, *Transportation Research Part E: Logistics and Transportation Review*, 167, p. 102906.

Zhang, K., Y. M. Nie (2018) Mitigating the impact of selfish routing: An optimal-ratio control scheme (ORCS) inspired by autonomous driving, *Transportation Research Part C: Emerging Technologies*, 87, pp. 75–90.

Zhang, K., Y. M. Nie (2021) To pool or not to pool: Equilibrium, pricing and regulation, *Transportation Research Part B: Methodological*, 151, pp. 59–90.

Zhang, W., S. Guhathakurta (2017) Parking spaces in the age of shared autonomous vehicles: How much parking will we need and where?, *Transportation Research Record*, 2651(1), pp. 80–91.

Zheng, Y., P. Meredith-Karam, A. Stewart, H. Kong, J. Zhao (2023) Impacts of congestion pricing on ride-hailing ridership: Evidence from Chicago, *Transportation Research Part A: Policy and Practice*, 170, p. 103639.

# Glossary

## List of abbreviations

The following abbreviations are used in this thesis:

| | |
|---|---|
| ALLM | Ajusted Lower Level Model |
| AMoD | Automated Mobility-on-Demand |
| AT | Automated Taxi |
| AV | Automated Vehicle |
| BO | Bayesian Optimisation |
| BPR | Bureau of Public Roads |
| CSC | China Scholarship Council |
| CT | Conventional Taxi |
| CV | Conventional Vehicle |
| DE | Differential Evolution |
| DHPC | Delft High Performance Computing |
| DTA | Dynamic Traffic Assignment |
| FIFO | First-In-First-Out |
| FSMVRP | Fleet Sizing and Mix Vehicle Routing Problem |
| GA | Genetic Algorithm |
| HV | Human-driven Vehicle |
| ITS | Intelligent Transportation Systems |
| LLM | Lower Level Model |
| MAA | Maximum Acceptable Approximation |
| MILP | Mixed-Integer Linear Programming |
| MINLP | Mixed-Integer Non-Linear Programming |
| MIP | Mixed-Integer Programming |
| MOZ | Mixed Operation Zone |
| OD | Origin Destination |
| ORAD | On-Road Automated Driving |

| | |
|---|---|
| PGA | Parallel Genetic Algorithm |
| PSO | Particle Swarm Optimisation |
| PV | Privately-owned Vehicle |
| SAE | Society of Automotive Engineers |
| SAV | Shared Automated Vehicle |
| SD | Standard Deviation |
| SO | System Optimum |
| SPM | System Profit Mode |
| TA | Traffic Assignment |
| TNC | Transportation Network Company |
| UE | User Equilibrium |
| ULM | Upper-Level Model |
| UPM | User Preference Mode |
| VOTT | Value of Travel Time |
| VRP | Vehicle Routing Problem |

# Summary

Shared automated vehicles (SAVs) are expected to revolutionise the transportation mobility system and contribute to the sustainable development of urban regions. The emergence of SAVs is anticipated to replace private cars, providing seamless door-to-door ride-hailing services that meet people's mobility needs. However, the introduction of SAVs is a gradual process, and numerous challenges may arise until the entire transportation system is fully automated. Firstly, there may be a gradual evolution of the infrastructure from traditional to intelligent transportation systems to better adapt to automated driving technology. Secondly, the mixed driving situation during this transition can be problematic. SAVs, functioning like moving robots, will follow SAV operator's route guidance to maximise system profits, whereas privately-owned human-driven vehicles (HVs) tend to operate selfishly to maximize individual utility. Their driving behaviours will interact with each other. Thirdly, when the entire city has transformed into a fully intelligent system with SAVs replacing all HVs, the future mobility demand for SAVs remains largely unknown, depending on supply side decisions, such as price and service quality.

The challenges mentioned above will significantly impact the decisions of SAV service providers. This thesis aims to help SAV service providers in making the most profitable planning decisions (fleet size, pricing, initial fleet distribution, service quality) and operational decisions (trip assignment, vehicle routing, parking, and relocation). These decisions will be tailored to various stages, ranging from a mixed driving environment to a fully automated driving environment, addressing each of the aforementioned challenges step by step.

We begin by studying the heterogeneous fleet sizing problem during a transitional stage in which certain city zones may be dedicated to automated vehicles (AVs), supported by a fully intelligent traffic management system. We propose a strategic flow-based vehicle routing model to determine the optimal fleet sizes of automated and conventional taxis, influenced by the gradually increasing coverage of the AVs-only dedicated area. Traffic congestion is considered through flow-dependent travel times. We test two taxi company service regimes: the User Preference Mode (UPM) and the

System Profit Mode (SPM). In the UPM, passengers can select their preferred vehicle type based on personal preferences. In the SPM, the taxi company manages vehicle assignments to maximize system profits.

To capture the mixed driving behaviours of centrally dispatched automated and conventional taxis, and privately-owned HVs and their interactions with infrastructure, we introduce a bi-level framework. The upper level decides the fleet sizes, while the lower level models the routing behaviours of centrally dispatched automated and conventional taxis (following system optimum) and HVs (following user equilibrium). We solve the proposed bi-level model using a parallel genetic algorithm with a tailored iterative algorithm to solve the lower-level routing model.

Next, we envision a fully automated scenario where SAVs replace all private cars, offering public on-demand mobility services. To model future mode choices between SAVs and active modes of transport, such as bicycles, across travellers in different income classes, we employ a binary logit model. Additionally, we develop a mixed-integer non-linear programming model that considers congestion effects and travellers' mode choice. We explore two types of trip acceptance mechanisms—mandatory and non-mandatory—that affect travellers' willingness to use SAV services. The complex non-linear nature of the model is addressed using reformulation techniques, outer-inner approximation methods, and a breakpoint determination algorithm.

Finally, we explore various pricing strategies: base fare plus distance-based fare, distance-based fare only, and income class-based fare. The developed models are non-linear systems that present significant challenges during the solving process. To address the model's complex nonlinearities, we developed three distinct solution algorithms, employing linearisation techniques, hybrid metaheuristic-based optimisation, and hybrid Bayesian optimisation-based methods.

In summary, this thesis provides mathematical models that enable SAV service providers to make optimal planning and operational decisions, facing the promising transition from a mixed driving environment to a fully automated environment. We have proposed various solution algorithms to address these models. By applying these methods to a case study in Delft, the Netherlands, we offer valuable managerial insights for SAV service operators.

# Samenvatting

Zelfrijdende deelauto's (SAV's: shared automated vehicles) worden gezien als een revolutie voor het mobiliteitssysteem en worden verwacht bij te dragen aan de verduurzaming van stedelijke regio's. Particuliere auto's zullen worden vervangen door SAV's die naadloze deur-tot-deur-ritdiensten leveren om aan de mobiliteitsbehoeften van mensen te voldoen. De introductie van SAV's is echter een geleidelijk proces en er kunnen vele uitdagingen ontstaan totdat het hele vervoerssysteem volledig geautomatiseerd is. Ten eerste kan een geleidelijke evolutie van de infrastructuur van traditionele naar intelligente transportsystemen nodig zijn om de geautomatiseerde rijtechnologie beter te kunnen accommoderen. Ten tweede kan de gemengde rijsituatie tijdens deze overgang problematisch zijn. SAV's, die functioneren als bewegende robots, zullen de winst van het systeem maximaliseren door de routeaanwijzingen van de SAV operator te volgen, terwijl door mensen bestuurde particuliere voertuigen (HV's: human-driven vehicles) geneigd zijn om zelfzuchtig te handelen om zo hun individuele nut te maximaliseren. Er zal interactie zijn tussen het rijgedrag van beide. Ten derde, wanneer de hele stad is getransformeerd naar een volledig intelligent systeem met SAV's die alle HV's vervangen, blijft de toekomstige mobiliteitsvraag voor SAV's grotendeels onbekend. Dit hangt sterk af van beslissingen aan de aanbodkant, zoals prijs en servicekwaliteit.

De genoemde uitdagingen zullen de beslissingen van SAV-dienstverleners aanzienlijk beïnvloeden. Het doel van dit proefschrift is om SAV-dienstverleners te helpen bij het maken van de meest winstgevende planningsbeslissingen (vlootgrootte, prijsstelling, initiële vlootverdeling, servicekwaliteit) en operationele beslissingen (rittoewijzing, voertuigroutering, parkeren en relocatie). Deze beslissingen zullen worden afgestemd op verschillende stadia, variërend van een gemengde rijomgeving tot een volledig geautomatiseerde rijomgeving, waarbij elk van de genoemde uitdagingen stap voor stap aan bod komen.

We beginnen met het bepalen van de grootte van heterogene vloten tijdens een overgangsfase waarin bepaalde stadszones mogelijk alleen toegankelijk zijn voor geautomatiseerde voertuigen (AV's: automated vehicles), ondersteund door een volledig

intelligent verkeersbeheersysteem. We formuleren een strategisch stroom gebaseerd voertuigrouteringsmodel om de optimale vlootgroottes van geautomatiseerde en conventionele taxi' te bepalen, beïnvloed door de geleidelijk toenemende dekking van het gebied dat alleen toegankelijk is voor AV's. Hierin worden verkeersopstoppingen meegenomen door de reistijd afhankelijk te maken van de verkeersstroom. We testen twee taxiservice-regimes: de modus waar gebruikersvoorkeur wordt meegenomen (UPM: User Preference Mode) en de modus die zich richt op systeemwinst (SPM: System Profit Mode). In de UPM kunnen passagiers hun voorkeur voor een voertuigtype aangeven op basis van persoonlijke voorkeuren. In de SPM beheert het taxibedrijf de toewijzing van voertuigen om de systeemwinst te maximaliseren.

Om het gemengde rijgedrag van centraal gestuurde geautomatiseerde en conventionele taxi's en particuliere HV's en hun interacties met de infrastructuur mee te nemen, introduceren we een kader op twee niveaus. Het bovenste niveau bepaalt de vlootgrootte, terwijl het lagere niveau het routegedrag modelleert van centraal gestuurde geautomatiseerde en conventionele taxi's (volgens het systeemoptimum) en HV's (volgens het gebruikersevenwicht). We lossen het voorgestelde tweelaagse model op met een parallel genetisch algoritme met een op maat gemaakt iteratief algoritme voor het oplossen van het routeringsmodel op het lagere niveau. Vervolgens bekijken we een volledig geautomatiseerd scenario waarbij alle particuliere voertuigen zijn vervangen door SAV's die openbare on-demand mobiliteitsdiensten aanbieden. Om toekomstige keuzes van vervoersmiddel tussen SAV's en actieve vervoerswijzen, zoals fietsen, te modelleren voor reizigers in verschillende inkomensklassen, gebruiken we een binair logit model. Daarnaast ontwikkelen we een niet-lineair gemengd geheeltallig programmeermodel dat rekening houdt met de effecten van verkeersopstoppingen en de keuze van de vervoerswijze van reizigers. We verkennen twee soorten acceptatiemechanismen - verplicht en niet-verplicht - die de bereidheid van reizigers om SAV-diensten te gebruiken beïnvloeden. De complexe niet-lineaire aard van het model wordt verholpen met behulp van herformuleringstechnieken, buiten-binnen benaderingsmethoden en een algoritme voor het bepalen van breekpunten.

Tot slot onderzoeken we verschillende prijsstrategieën: starttarief plus kilometertarief, alleen kilometertarief en inkomensklasse-gebaseerd tarief. De ontwikkelde modellen zijn niet-lineaire systemen die erg uitdagend zijn om op te lossen. Om de complexe niet-lineariteiten van het model aan te pakken, hebben we drie verschillende oplossingsalgoritmen ontwikkeld, waarbij linearisatietechnieken, hybride metaheuristiek-gebaseerde optimalisatie en hybride Bayesiaanse optimalisatiemethoden worden gebruikt.

Samenvattend worden er in dit proefschrift wiskundige modellen geformuleerd die SAV-dienstverleners in staat stellen optimale plannings- en operationele beslissingen te

maken tijdens de veelbelovende overgang van een gemengde rijomgeving naar een volledig geautomatiseerde omgeving. Verschillende oplossingsalgoritmen worden voorgesteld om deze modellen op te lossen. Door deze methoden toe te passen op een casestudy in Delft (Nederland), bieden we waardevolle managementinzichten voor exploitanten van SAV-services.

# About the author

Qiaochu Fan was born on June 14, 1994, in China. She obtained her Bachelor of Science in Transportation Planning and Engineering from the Mao Yisheng Honour School at Southwest Jiaotong University in Chengdu. She holds a Master of Engineering in Information Science and Digital Society from École Centrale Marseille in France and a Master of Science in Transportation Planning and Management from Southwest Jiaotong University.

In September 2019, Qiaochu Fan began her PhD in the Discrete Mathematics & Optimisation research group at Delft University of Technology in the Netherlands. Under the supervision of Dr. J. Theresia van Essen and Dr. Gonçalo H. A. Correia, her research focuses on operations research, choice modelling, integer programming, and metaheuristics.

# Publications

1. **Fan, Q.**, van Essen, J. T., & Correia, G. H. (2024). A bi-level framework for heterogeneous fleet sizing of ride-hailing services considering an approximated mixed equilibrium between automated and non-automated traffic. *European Journal of Operational Research*, 315(3), 879-898.

2. Ni, X., Hu, W., **Fan, Q.**, Cui, Y., & Qi, C. (2024). A Q-learning based multi-strategy integrated artificial bee colony algorithm with application in unmanned vehicle path planning. *Expert Systems with Applications*, 236, 121303.

3. **Fan, Q.**, van Essen, J. T., & Correia, G. H. (2023). Optimising fleet sizing and management of shared automated vehicle (SAV) services: A mixed-integer programming approach integrating endogenous demand, congestion effects, and accept/reject mechanism impacts. *Transportation Research Part C: Emerging Technologies*, 157, 104398.

4. **Fan, Q.**, van Essen, J. T., & Correia, G. H. (2022). Heterogeneous fleet sizing for on-demand transport in mixed automated and non-automated urban areas. *Transportation Research Procedia*, 62, 163-170.

5. **Fan, Q.**, van Essen, J. T., & Correia, G. H. (Under review). Solution methods for pricing and fleet management in shared automated vehicle services considering supply-demand dynamics, congestion, and income heterogeneity.

# TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 400 titles, see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Fan, Q., *Fleet Management Optimisation for Ride-hailing Services: from mixed traffic to fully automated environments,* T2025/4, April 2025, TRAIL Thesis Series, the Netherlands

Hagen, L. van der, *Machine Learning for Time Slot Management in Grocery Delivery,* T2025/3, March 2025, TRAIL Thesis Series, the Netherlands

Schilt, I.M. van, *Reconstructing Illicit Supply Chains with Sparse Data: a simulation approach,* T2025/2, January 2025, TRAIL Thesis Series, the Netherlands

Ruijter, A.J.F. de, *Two-Sided Dynamics in Ridesourcing Markets,* T2025/1, January 2025, TRAIL Thesis Series, the Netherlands

Fang, P., *Development of an Effective Modelling Method for the Local Mechanical Analysis of Submarine Power Cables,* T2024/17, December 2024, TRAIL Thesis Series, the Netherlands

Zattoni Scroccaro, P., *Inverse Optimization Theory and Applications to Routing Problems,* T2024/16, October 2024, TRAIL Thesis Series, the Netherlands

Kapousizis, G., *Smart Connected Bicycles: User acceptance and experience, willingness to pay and road safety implications,* T2024/15, November 2024, TRAIL Thesis Series, the Netherlands

Lyu, X., *Collaboration for Resilient and Decarbonized Maritime and Port Operations,* T2024/14, November 2024, TRAIL Thesis Series, the Netherlands

Nicolet, A., *Choice-Driven Methods for Decision-Making in Intermodal Transport: Behavioral heterogeneity and supply-demand interactions,* T2024/13, November 2024,

TRAIL Thesis Series, the Netherlands

Kougiatsos, N., *Safe and Resilient Control for Marine Power and Propulsion Plants,* T2024/12, November 2024, TRAIL Thesis Series, the Netherlands

Uijtdewilligen, T., *Road Safety of Cyclists in Dutch Cities,* T2024/11, November 2024, TRAIL Thesis Series, the Netherlands

Liu, X., *Distributed and Learning-based Model Predictive Control for Urban Rail Transit Networks,* T2024/10, October 2024, TRAIL Thesis Series, the Netherlands

Clercq, G. K. de, *On the Mobility Effects of Future Transport Modes,* T2024/9, October 2024, TRAIL Thesis Series, the Netherlands

Dreischerf, A.J., *From Caveats to Catalyst: Accelerating urban freight transport sustainability through public initiatives,* T2024/8, September 2024, TRAIL Thesis Series, the Netherlands

Zohoori, B., *Model-based Risk Analysis of Supply Chains for Supporting Resilience,* T2024/7, October 2024, TRAIL Thesis Series, the Netherlands

Poelman, M.C., *Predictive Traffic Signal Control under Uncertainty: Analyzing and Reducing the Impact of Prediction Errors,* T2024/6, October 2024, TRAIL Thesis Series, the Netherlands

Berge, S.H., *Cycling in the age of automation: Enhancing cyclist interaction with automated vehicles through human-machine interfaces,* T2024/5, September 2024, TRAIL Thesis Series, the Netherlands

Wu, K., *Decision-Making and Coordination in Green Supply Chains with Asymmetric Information,* T2024/4, July 2024, TRAIL Thesis Series, the Netherlands

Wijnen, W., *Road Safety and Welfare,* T2024/3, May 2024, TRAIL Thesis Series, the Netherlands

Caiati, V., *Understanding and Modelling Individual Preferences for Mobility as a Service,* T2024/2, March 2024, TRAIL Thesis Series, the Netherlands

Vos, J., *Drivers' Behaviour on Freeway Curve Approach,* T2024/1, February 2024, TRAIL Thesis Series, the Netherlands

Geržinič, N., *The Impact of Public Transport Disruptors on Travel Behaviour,* T2023/20, December 2023, TRAIL Thesis Series, the Netherlands

Dubey, S., *A Flexible Behavioral Framework to Model Mobility-on-Demand Service Choice Preference,* T2023/19, November 2023, TRAIL Thesis Series, the Netherlands

Sharma, S., *On-trip Behavior of Truck Drivers on Freeways: New mathematical models and control methods,* T2023/18, October 2023, TRAIL Thesis Series, the Nether-

lands

Ashkrof, P., *Supply-side Behavioural Dynamics and Operations of Ride-sourcing Platforms,* T2023/17, October 2023, TRAIL Thesis Series, the Netherlands

Sun, D., *Multi-level and Learning-based Model Predictive Control for Traffic Management,* T2023/16, October 2023, TRAIL Thesis Series, the Netherlands

Brederode, L.J.N., *Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems,* T2023/15, October 2023, TRAIL Thesis Series, the Netherlands

Hernandez, J.I., *Data-driven Methods to study Individual Choice Behaviour: with applications to discrete choice experiments and Participatory Value Evaluation experiments,* T2023/14, October 2023, TRAIL Thesis Series, the Netherlands