TU Delft

Delft University of Technology

A bi-level framework for heterogeneous fleet sizing of ride-hailing services considering an approximated mixed equilibrium between automated and non-automated traffic

Fan, Qiaochu; van Essen, J. Theresia; Correia, Gonçalo H.A.

DOI
10.1016/j.ejor.2024.01.017

Publication date
2024

Document Version
Final published version

Published in
European Journal of Operational Research

Citation (APA)
Fan, Q., van Essen, J. T., & Correia, G. H. A. (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. https://doi.org/10.1016/j.ejor.2024.01.017

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

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

European Journal of Operational Research 315 (2024) 879–898

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Discrete Optimization

A bi-level framework for heterogeneous fleet sizing of ride-hailing services considering an approximated mixed equilibrium between automated and non-automated traffic

Qiaochu Fan [a,*], J. Theresia van Essen [a], Gonçalo H.A. Correia [b]

[a] Delft Institute of Applied Mathematics, Delft University of Technology, 2628 CD, Delft, The Netherlands
[b] Department of Transport & Planning, Delft University of Technology, 2628 CN, Delft, The Netherlands

ARTICLE INFO

Keywords:
Routing
Fleet sizing
Approximated mixed equilibrium
AVs-only zone
Ride-hailing services

ABSTRACT

Ride-hailing companies will face the emergence and gradual expansion of AVs-only zones in urban areas where only automated vehicles (AVs) are allowed to circulate. When owning a mixed fleet (automated and conventional taxis), a ride-hailing company has to determine the optimal fleet size as a function of the gradually expanding coverage of AVs-only zones while taking into account interactions with privately-owned human-driven vehicles. To model this problem, we propose a bi-level framework in which the lower level captures the mixed routing behaviour of the vehicles and the endogenous traffic congestion, and the upper level determines fleet sizes to maximise profit. A parallel genetic algorithm is introduced to solve this bi-level framework, which is embedded with a tailored algorithm for solving the lower-level model. Numerical experiments are conducted on instances based on a small network and the network of the city of Delft, The Netherlands, to demonstrate the performance of the proposed solution method and investigate the impacts of AVs-only zones on traffic and ride-hailing operations. Results indicate that the fleet size of automated taxis increases nonlinearly with the expansion of the AVs-only zone while that of conventional taxis decreases as demand shifts from human-driven vehicles to automated taxis. The fleet size decision depends heavily on the fleet's cost structure, the location and the distribution of parking depots. Furthermore, the existence of an AVs-only zone leads to detours for human-driven vehicles in the early stages, but it will bring major benefits by reducing congestion as its size increases.

1. Introduction

Uber's establishment in 2009 marked the beginning of the ride-hailing industry. Since then, an increasing number of ride-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 (Vega-Gonzalo et al., 2023). 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, ride-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., 2022), and offering continuous, high-quality door-to-door trip services (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 (Olia et al., 2018; Yang et al., 2016). 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; Conceição et al., 2021; Madadi et al.,

* Corresponding author.
E-mail address: q.fan-1@tudelft.nl (Q. Fan).

https://doi.org/10.1016/j.ejor.2024.01.017
Received 1 June 2023; Accepted 16 January 2024
Available online 18 January 2024
0377-2217/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2020) to AVs. These areas, which we will designate in this paper 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 (Mo et al., 2022; Scherr et al., 2019; Yang et al., 2020). 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, Hao, et al., 2021). But less attention has been devoted to dynamic interactions between road users and the infrastructure, resulting in endogenous traffic congestion.

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 paper 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 their explicit modelling. The key challenge of this paper 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 (Liang et al., 2020; Stoiber et al., 2019); 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 paper 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 ride-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.

The remaining sections of this paper are structured as follows. The literature on fleet sizing problems, vehicle routing problems (VRP) and traffic assignment (TA) are reviewed in Section 2. Section 3 presents the mathematical model of the proposed bi-level framework. Then, in Section 4, a detailed description of solution methods for the lower level and the entire problem is provided. In Section 5, a small toy network case study and a quasi-real case study of the city of Delft in the Netherlands are carried out to demonstrate the effectiveness of the proposed framework and to evaluate the impact of AVs-only zones on all the traffic participants. Conclusions and future outlook are given in Section 6.

2. Literature review

2.1. Fleet sizing problem for ride-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 (Balac et al., 2020; Hiermann et al., 2016), 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. Guo, Hao, et al. (2021) 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. Thus, 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 paper.

Mainly three types of modelling techniques have been used to tackle fleet sizing problems: simulation-based techniques (Fagnant & Kockelman, 2018; Wang et al., 2022; Yi & Smart, 2021), optimisation-based techniques (Allahviranloo & Chow, 2019; Balac et al., 2020; Guo, Hao, et al., 2021), 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 MILP model (Balac et al., 2020; Koç et al., 2016; Santos & Correia, 2021), or a bi-level model (Allahviranloo & Chow, 2019), solved by exact methods (Balac et al., 2020; Fan et al., 2022; Santos & Correia, 2021), or heuristic methods (Brandão, 2009; Koç et al., 2016; Renaud & Boctor, 2002), 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 and 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 and 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 paper 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.

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 paper 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 and 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 and 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 and 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; Kashmiri & Lo, 2022; Ke & Qian, 2023; Zhang et al., 2022; Zhang & Nie, 2018), day-to-day dynamic systems (Li et al., 2018; Liang et al., 2023), and dynamic scenarios (Guo, Ban, & Aziz, 2021; Hoang et al., 2023; Mansourianfar et al., 2022, 2021), 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 paper, 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 proposed 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. Problem formulation

The proposed bi-level framework is presented in this section as a bi-level MILP model. In Section 3.1, we first introduce the problem. Then, we propose the mathematical formulation of the upper level and the lower level in Sections 3.2 and 3.3, respectively.

3.1. Problem description and modelling framework

The demand of travellers heading from origins to destinations triggers the need to plan the operation of ride-hailing services and vehicle movements on the road network. The model structure that is supposed to solve the problem is presented in Fig. 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 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 ride-hailing service. Assuming that travellers who use ride-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, 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 TNC 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 1).

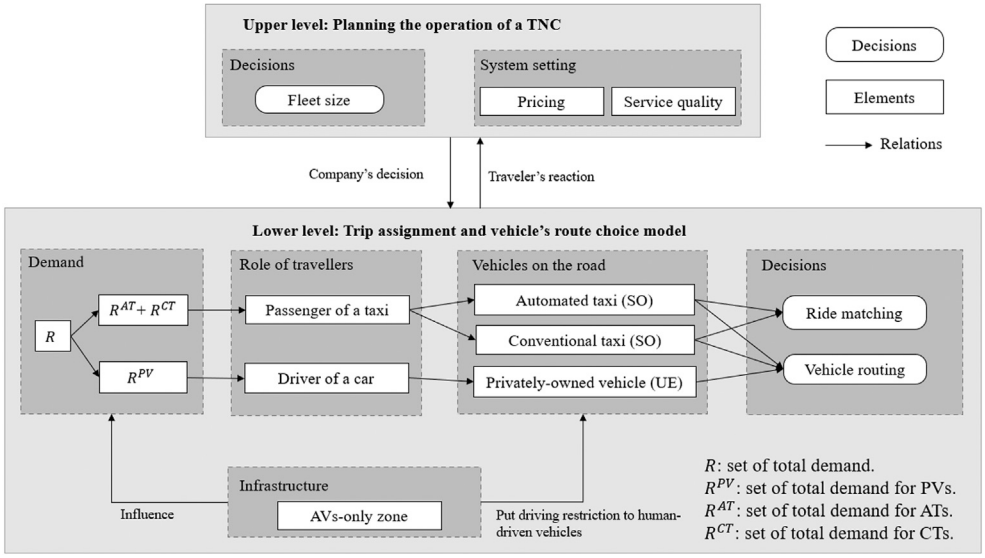

Fig. 1. Decisions, elements and their relations in the bi-level optimisation problem.

Table 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.

3.2. Upper level: Planning for the TNC

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

$$\begin{aligned}[\mathbf{ULM}]\ \max Z = & \sum_{m\in\{CT,AT\}}\sum_{r\in R^m}\left(p^0 P^r + p^m P^r sd^r\right) - s\cdot cp\cdot V^{CT} \\ & - \sum_{m\in\{CT,AT\}} cf^m V^m \\ & - \sum_{m\in\{CT,AT\}} co^m \left(\sum_{(i_{t_1},j_{t_2})\in G^m} l_{ij} F^m_{i_{t_1}j_{t_2}}\right) \\ & - cd\sum_{m\in\{CT,AT\}}\sum_{r\in R^m}\left(\sum_{t\in T} tE^{rt} - a^r n^r - st^r n^r\right)\end{aligned} \tag{1}$$

Subject to:

$$E^{rt}, F^m_{i_{t_1}j_{t_2}} \in \arg\min\{\text{Objective function (5)–(7)} : \text{Constraints (8)–(28)}\} \tag{2}$$

$$lb^m \le V^m \le ub^m, \forall m \in \{CT, AT\} \tag{3}$$

$$\alpha n^r \le P^r \le n^r, \forall r \in R^m, m \in \{CT, AT\} \tag{4}$$

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 Eq. (2). Constraints (3) impose an upper bound and lower bound on the total fleet size of CTs and ATs which is explained in Section 4.2. Constraints (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.

Table 2
Notation.

Notation	Description
Set	
M	$= \{CT, AT, PV\}$, set of vehicle types.
T	$= \{0, \dots, t, \dots, s\}$, set of time instants in the operation period.
N	$= \{1, \dots, i, \dots\}$, set of nodes.
L	$= \{\dots, (i,j), \dots\}$, set of road links between nodes in set N.
G	$= \{\dots, (i_{t_1}, j_{t_2}), \dots\}$, set of links in the time-space network.
R^m	$= \{1, \dots, r, \dots\}$, 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, 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.
Π^r	$= \{1, \dots, \pi, \dots\}$, set of paths of group of trips $r \in R^{PV}$.
Parameters	
p^0	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.
α	Minimum service rate for orders.
lb^m, ub^m	Lower bound and upper bound of taxi's fleet size of type $m \in \{CT, AT\}$.
ω	Calibrated weighting coefficient to combine two objective functions into one.
λ	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.
Variables	
P^r	Integer variable representing the total number of served trips from group r, where $r \in R^m, m \in \{CT, AT\}$.
$PF_{i_{t_1} j_{t_2}}^r$	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 $(i_{t_1}, j_{t_2}) \in G^m, a^r \le t_1 < t_2 \le b^r$. If $t_1 = a^r$, then $i = o^r$.
$PF_{i_{t_1} j_{t_2}}^{r\pi}$	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 $\delta_{ij}^{r\pi} = 1$, $a^r \le t_1 < t_2 \le b^r$.
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 $t \in T$.
$F_{i_{t_1} j_{t_2}}^m$	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 $(i_{t_1}, j_{t_2}) \in G^m$.
$W_{i_t}^m$	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. 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 OD 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 and Correia (2019) the concept of approximated DUE 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 paper 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 OD 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 objectives. 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 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 OD 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 OD 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 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

$$[\mathbf{LLM}] \qquad \min J = \omega J^T + (1-\omega) J^P \tag{5}$$

where

$$J^T = \sum_{m\in\{CT,AT\}} co^m \left(\sum_{\left(i_{t_1}, j_{t_2}\right)\in G^m} l_{ij} F^m_{i_{t_1} j_{t_2}} \right) + cd \sum_{r\in R^m} \sum_{m\in\{CT,AT\}} \left(\sum_{t\in T} t E^{rt} - a^r P^r - st^r P^r \right) \tag{6}$$

$$J^P = \lambda \sum_{r\in R^{PV}} \frac{K^r}{M^r} + \sum_{\pi\in\Pi^r, r\in R^{PV}} \sum_{\left(i_{t_1}, d^r_t\right)\in G^{PV}} PF^{r\pi}_{i_{t_1} d^r_t} \left(co^{PV} D^{r\pi} + ct(t-a^r) \right) \tag{7}$$

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 λ that gives absolute priority to this term. A detailed description of how to determine the value of λ can be found in Appendix A.2. As previously stated, we use the weighted-sum method to combine J^T and J^P into one single objective function (weight ω). The objective function is constrained by the following:

Constraints for taxis:

$$P^r = \sum_{j_t \mid \left(o^r_{a^r}, j_t\right)\in G^m} PF^r_{o^r_{a^r} j_t}, \forall r \in R^m, m \in \{CT, AT\} \tag{8}$$

$$P^r = \sum_{t\in T \mid a^r + st^r \le t \le b^r} E^{rt}, \forall r \in R^m, m \in \{CT, AT\} \tag{9}$$

$$E^{rt} = \sum_{i_{t_1} \mid \left(i_{t_1}, d^r_t\right)\in G^m} PF^r_{i_{t_1} d^r_t}, \forall r \in R^m, m \in \{CT, AT\}, t \in T \tag{10}$$

$$\sum_{j_{t_2} \mid \left(d^r_{t_1}, j_{t_2}\right)\in G^m} PF^r_{d^r_{t_1} j_{t_2}} = 0, \forall r \in R^m, m \in \{CT, AT\}, t_1 \in T, a^r + st^r \le t_1 \le b^r \tag{11}$$

$$\sum_{i_{t_1} \mid \left(i_{t_1}, o^r_{t_2}\right)\in G^m} PF^r_{i_{t_1} o^r_{t_2}} = 0, \forall r \in R^m, m \in \{CT, AT\}, t_2 \in T, a^r \le t_2 \le b^r \tag{12}$$

$$\sum_{j_{t_0} \mid \left(j_{t_0}, i_{t_1}\right)\in G^m} PF^r_{j_{t_0} i_{t_1}} = \sum_{j_{t_2} \mid \left(i_{t_1}, j_{t_2}\right)\in G^m} PF^r_{i_{t_1} j_{t_2}}, \forall r \in R^m, m \in \{CT, AT\}, t_1 \in T, a^r < t_1 < b^r, i \in N^m, i \ne o^r, i \ne d^r \tag{13}$$

$$\sum_{r\in R^m} PF^r_{i_{t_1} j_{t_2}} \le F^m_{i_{t_1} j_{t_2}}, \forall \left(i_{t_1}, j_{t_2}\right) \in G^m, m \in \{CT, AT\} \tag{14}$$

$$\sum_{(i_0, j_t)\in G^m} F^m_{i_0 j_t} + \sum_{i\in N^m_P} W^m_{i_0} = V^m, \forall m \in \{CT, AT\} \tag{15}$$

$$\sum_{j_{t_1} \mid \left(j_{t_1}, i_t\right)\in G^m, t_1 < t} F^m_{j_{t_1} i_t} + W^m_{i_{t-1}} = \sum_{j_{t_2} \mid \left(i_t, j_{t_2}\right)\in G^m, t < t_2} F^m_{i_t j_{t_2}} + W^m_{i_t}, \forall t \in T, 0 < t < s, i \in N^m_P, m \in \{CT, AT\} \tag{16}$$

$$\sum_{j_{t_1} \mid \left(j_{t_1}, i_t\right)\in G^m, t_1 < t} F^m_{j_{t_1} i_t} = \sum_{j_{t_2} \mid \left(i_t, j_{t_2}\right)\in G^m, t < t_2} F^m_{i_t j_{t_2}}, \forall t \in T, 0 < t < s, i \in N^m \setminus N^m_P, m \in \{CT, AT\} \tag{17}$$

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 (8)–(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 (11) and (12) guarantee that the passenger flows start at the origin node and end at the destination node. Constraints (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 (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 (15) guarantee that the total number of taxis circulating on road link (i,j) or parking at depot $i \in N_p^m$ 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 (16) and (17) respectively.

Constraints for PVs:

$$\sum_{\pi \in \Pi^r} F^{r\pi} = n^r, \forall r \in R^{PV} \tag{18}$$

$$F^{r\pi} = \sum_{j_{t_2} | \left(o^r_{a^r}, j_{t_2}\right) \in G^{PV}, \delta^{r\pi}_{o^r j} = 1} PF^{r\pi}_{o^r_{a^r}, j_{t_2}}, \forall \pi \in \Pi^r, r \in R^{PV} \tag{19}$$

$$F^{r\pi} = \sum_{\left(i_{t_1}, d^r t_2\right) \in G^{PV}, \delta^{r\pi}_{id^r} = 1} PF^{r\pi}_{i_{t_1}, d^r t_2}, \forall \pi \in \Pi^r, r \in R^{PV} \tag{20}$$

$$\sum_{j_{t_0} | \left(j_{t_0}, i_{t_1}\right) \in G^{PV}, \delta^{r\pi}_{ji} = 1} PF^{r\pi}_{j_{t_0} i_{t_1}} = \sum_{j_{t_2} | \left(i_{t_1}, j_{t_2}\right) \in G^{PV}, \delta^{r\pi}_{ij} = 1} PF^{r\pi}_{i_{t_1}, j_{t_2}}, \forall \pi \in \Pi^r, r \in R^{PV}, t_1 \in T, a^r < t_1 < b^r, i \in N^{PV}, i \neq o^r, i \neq d^r \tag{21}$$

$$F^{PV}_{i_{t_1} j_{t_2}} = \sum_{\pi \in \Pi^r, r \in R^{PV}} PF^{r\pi}_{i_{t_1} j_{t_2}}, \forall \left(i_{t_1}, j_{t_2}\right) \in G^{PV} \tag{22}$$

$$A^{r\pi}_t \geq \frac{\sum_{i_{t_1} | \left(i_{t_1}, d^r_t\right) \in G^{PV}} PF^{r\pi}_{i_{t_1}, d^r_t}}{n^r}, \forall \pi \in \Pi^r, r \in R^{PV}, t \in T \tag{23}$$

$$K^{r\pi} = \sum_{t \in T} A^{r\pi}_t \left(co^{PV} D^{r\pi} + ct(t - a^r)\right), \forall \pi \in \Pi^r, r \in R^{PV} \tag{24}$$

$$K^r \geq K^{r\pi}, \forall \pi \in \Pi^r, r \in R^{PV} \tag{25}$$

Constraints (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 (19) and (20). Constraints (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 (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 $D^{r\pi} = \sum_{(i,j) \in L^{PV}} l_{ij} \delta^{r\pi}_{ij}$. Constraints (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 (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 (25).

Constraints for traffic congestion:

$$\sum_{m \in M} F^m_{i_{t_1} j_{t_2}} \leq \left\lfloor C_{i_{t_1} j_{t_2}} \right\rfloor X_{i_{t_1} j_{t_2}}, \forall \left(i_{t_1}, j_{t_2}\right) \in G \tag{26}$$

$$\sum_{t_2 | \left(i_{t_1}, j_{t_2}\right) \in G} X_{i_{t_1} j_{t_2}} \leq 1, \forall (i,j) \in L, t_1 \in T \tag{27}$$

$$t_1 + \sum_{t \in T} X_{i_{t_1} j_t} (t - t_1) \leq t_2 + \sum_{t \in T} X_{i_{t_2} j_t} (t - t_2) + M \left(1 - \sum_{t \in T} X_{i_{t_2} j_t}\right), \forall t_1, t_2 \in T, t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij}, (i,j) \in L \tag{28}$$

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 $t^{\min}_{ij}$, which is

$$C_{i_{t_1} j_{t_2}} = (t_2 - t_1) Q_{ij} \left(\frac{1}{a}\left(\frac{t_2 - t_1}{t^{\min}_{ij}} - 1\right)\right)^{\frac{1}{b}}. \tag{29}$$

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 (26) and (27). Constraints (26) impose an additional requirement that the total flow on road link (i,j) never exceeds its spatial link capacity. Constraints (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 $t_1 < t_2 \leq t_1 + t^{\max}_{ij} - t^{\min}_{ij}$. Otherwise, if $t_2 > t_1 + t^{\max}_{ij} - t^{\min}_{ij}$, rewritten as $t_2 + t^{\min}_{ij} > t_1 + t^{\max}_{ij}$, 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.

4. Solution method

In this section, we first propose a two-stage solution method to solve the LLM in Section 4.1. Then, in Section 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.

4.1. Solution method for the LLM

One question remains to be tackled before we can solve the proposed LLM in Section 3.3: how to generate the set of paths Π^r for each group of trips $r \in R^{PV}$. The set of paths Π^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 OD 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

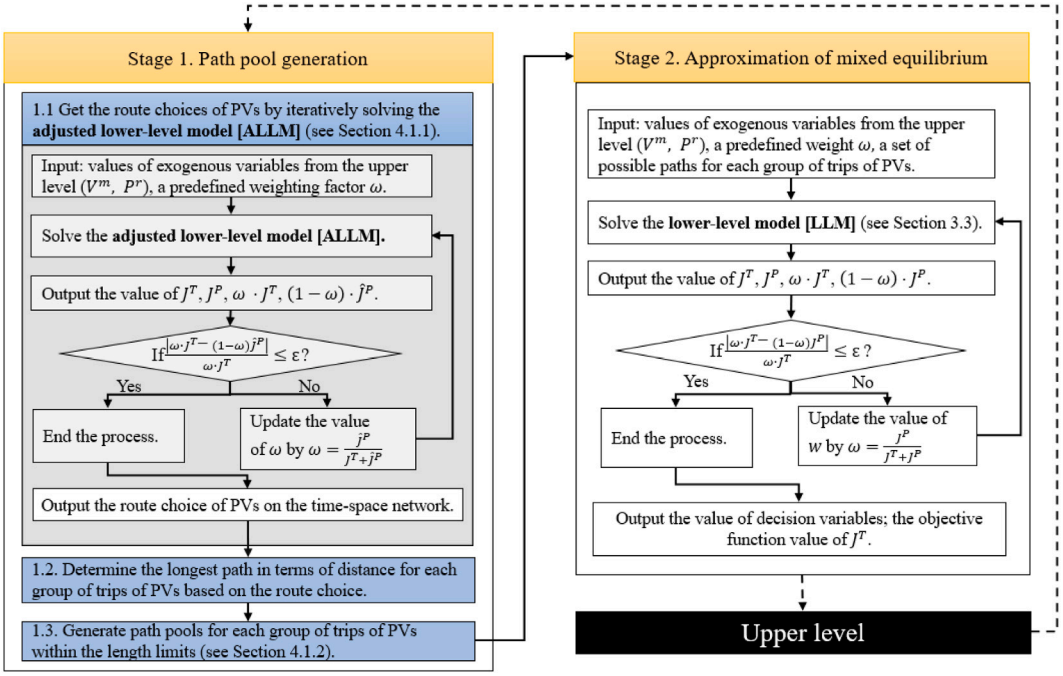

Fig. 2. Framework of the lower-level solution method.

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.

To solve the LLM, we propose the two-stage solution method depicted in Fig. 2. At Stage 1, we propose a method for generating a path pool for each OD 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 OD pair is identified via iteratively solving an additional MILP model which is adjusted from the proposed LLM in Section 3.3. The mathematical formulation of this model is presented in Section 4.1.1. The procedure is embedded with the weight determination algorithm described in Appendix A.1. The path enumeration with length limits is presented in Section 4.1.2. 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.

4.1.1. 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 OD 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.

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. Following is the formulation of the ALLM.

Objective function

$$[\textbf{ALLM}] \qquad \min J = \omega J^T + (1-\omega)\hat{J}^P \tag{30}$$

where

$$\hat{J}^P = \lambda \sum_{r\in R^{PV}} \frac{m^r}{st^r} + \sum_{r\in R^{PV}} \left(\sum_{\left(i_{t_1}, d_t^r\right)\in G^{PV}} tPF^r_{i_{t_1},d_t^r} - a^r n^r \right) \tag{31}$$

The objective function is updated to Eq. (30), with J^T remaining unchanged from Eq. (6) and $\hat{J}^P$ represented by Eq. (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 (30) is subject to Constraints (8)–(17), (26)–(28), and (32)–(37).

$$\sum_{j_t|\left(o^r_{a^r},j_t\right)\in G^{PV}} PF^r_{o^r_{a^r}j_t} = n^r, \forall r \in R^{PV} \tag{32}$$

$$\sum_{\left(i_{t_1},d^r_{t_2}\right)\in G^{PV}} PF^r_{i_{t_1}d^r_{t_2}} = n^r, \forall r \in R^{PV} \tag{33}$$

$$\sum_{j_{t_0}|\left(j_{t_0},i_{t_1}\right)\in G^{PV}} PF^r_{j_{t_0}i_{t_1}} = \sum_{j_{t_2}|\left(i_{t_1},j_{t_2}\right)\in G^{PV}} PF^r_{i_{t_1}j_{t_2}},$$
$$\forall r \in R^{PV}, t_1 \in T, t_0 < t_1 < t_2, i \in N^{PV}, i \neq o^r, i \neq d^r \tag{34}$$

$$A^r_t \ge \frac{\sum_{i_{t_1}|\left(i_{t_1},d^r_t\right)\in G^{PV}} PF^r_{i_{t_1},d^r_t}}{n^r}, \forall r \in R^{PV}, t \in T, a^r \le t \le b^r \tag{35}$$

Table 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 $r \in R^{PV}$.

$$m^r \geq tA_t^r - a^r, \forall r \in R^{PV}, t \in T \tag{36}$$

$$F^{PV}_{i_{t_1} j_{t_2}} = \sum_{r \in R^{PV}} PF^r_{i_{t_1} j_{t_2}}, \forall \left(i_{t_1}, j_{t_2}\right) \in G^{PV} \tag{37}$$

Constraints (32) and (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 (34). The arrival times of trips in group $r \in R^{PV}$ are specified in constraints (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 (36). The movement of PVs is identical to the movement of travellers within the cars. Constraints (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 OD pair can be identified.

4.1.2. 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 (Chondrogiannis et al., 2020; Liu et al., 2017). This method consists of removing paths whose similarity to any selected paths exceeds a predetermined threshold θ. Schnabel and Löhse (1997) proposed that the paths are not considered separate if they overlap more than 50%. In this paper, 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 1. By reducing the number of possible paths in the path pools, the number of variables and constraints in the LLM are reduced.

Algorithm 1 Similarity-based path pool size reduction procedure

Input: similarity threshold θ, longest distance thresholds ld^r for $r \in R^{PV}$.
Output: $PathPoolUpdated$ (a list).

```
Initialise empty lists PathPool := [[ ] for r ∈ R^PV], PathPoolUpdated := [[ ] for r ∈ 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 θ' between path1 and path2.
            if θ' > θ 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
```

4.2. Parallel genetic algorithm (PGA)

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 non-linear 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

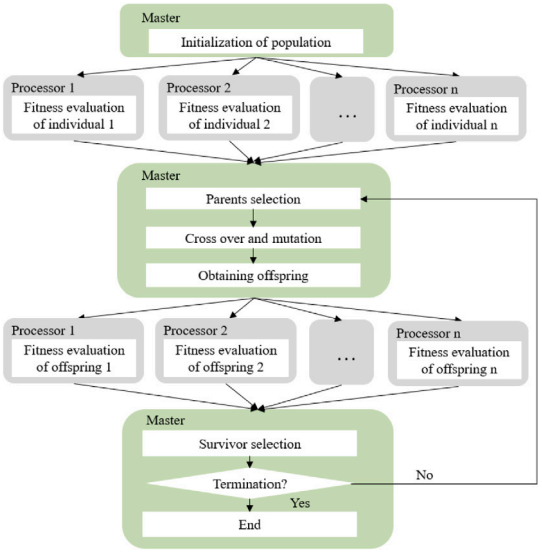

Fig. 3. Structure of the parallel genetic algorithm (PGA).

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 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 paper, 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.

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 parallelled 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. Fig. 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 paper, 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. Fig. 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 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.

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 paper, 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

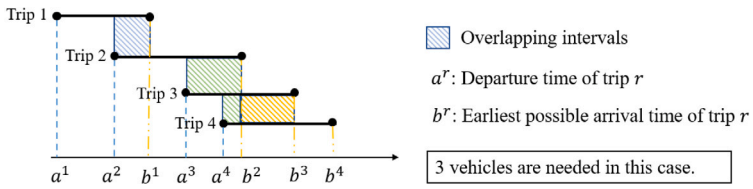

Fig. 4. Illustration of finding the minimum number of taxis to serve four trips.

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.

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 5.1. Then, in Section 5.2, we apply the proposed bi-level model to a quasi-real case study representing the city of Delft, in the Netherlands.

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 Fig. 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 min) and 4 time steps (10 min), 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.

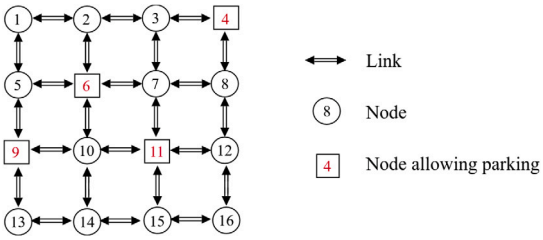

Fig. 5. Illustration of the small toy network.

Six groups of trips are considered, with the trip information shown in Table 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 (Bösch et al., 2018; Kouwenhoven et al., 2014).

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 ω as 0.5. The algorithm terminates when the relative difference between the contributed values is smaller than 5%.

Table 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

Table 5
Results of the base scenario in the small toy network.

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

Table 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] PV (UE): [1-2-3-7-11-10], [1-2-6-10]	34, 53, 53 48, 92	6, 6, 6 10, 6	7, 4, 4 7, 7
11-6	ALLM	Taxi (SO): [11-10-6], [11-7-6] PV (UE): [11-12-8-7-6], [11-15-14-10-6], [11-7-6]	53, 77 53, 53, 24	4, 4 8, 8, 4	2, 4 4, 4, 4
5-7	ALLM	Taxi (SO): [5-6-7], [5-1-2-6-7], [5-1-2-3-7] PV (UE): [5-6-7], [5-9-10-11-7]	59, 8, 53 67, 53	4, 8, 8 4, 8	4, 6, 6 4, 4

5.1.1. Computational results at stage 1: Path pool generation

The computational results are shown in Table 5, demonstrating that three iterations are needed to satisfy the convergence criterion and accurately determine the value of ω 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 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 10 km, 8 km, and 8 km, 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.

5.1.2. 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 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 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 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.

5.1.3. 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 9. Similar patterns happen for the other OD 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.

5.2. Quasi-real case study of the city of Delft, in the Netherlands

5.2.1. 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 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

Table 7
Final results of the base scenario in the small toy network.

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

Table 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] PV (UE): [1-2-6-10], [1-5-6-10]	87, 53 39, 101	6, 6 6, 6	6, 5 6, 7
11-6	LLM	Taxi (SO): [11-12-8-7-6], [11-7-6] PV (UE): [11-7-6], [11-10-6]	8, 122 4, 126	8, 4 4, 4	4, 4 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] PV (UE): [5-6-7]	53, 6, 8, 53 120	8, 4, 12, 8 4	4, 4, 7, 5 4

Table 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

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 50 km/h and 70 km/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.

Fig. 6 initially depicts the conventional road network, where there is no AVs-only zone. The AVs-only zone is then gradually expanded, covering 25%, 50%, 75%, and 100% of the links. To expand the AVs-only zone, we initially define it in areas characterised by frequent traffic congestion, such as the city centre, train 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 paper. 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 (1) subject to constraints (3), (4), (8)–(17) and (26)–(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 OD 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 10. For simplification purposes, the minimum service rate α 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 α will be studied in future research. The appendix contains the parameter tuning for the similarity threshold and population size.

5.2.2. 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%. Fig. 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

Fig. 6. Road networks of Delft with different AVs-only zone size: 0%, 25%, 50%, 75%, and 100%.

Table 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	0.15 euros/min	Maximum number of generations	100
p^0	3 euros/trip	Maximum number of iterations with	20
$p^m, m \in \{CT, AT\}$	2.55, 2.3 euros/km	no change for the best solution	
Minimum service rate (α)	1	Maximum number of iterations with	10
Similarity threshold (θ)	80%	no change for the quality of the top	
Relative difference threshold (ϵ)	5%	five fittest individuals	
		Maximum number of iterations with	10
Soft time limit	30 mins	no change for the quality of the top	
Hard time limit	60 mins	five fittest individuals	
MILP gap limit	2%		

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 23.7 h, 13.6 h, 4 h, and 6.7 h, 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.

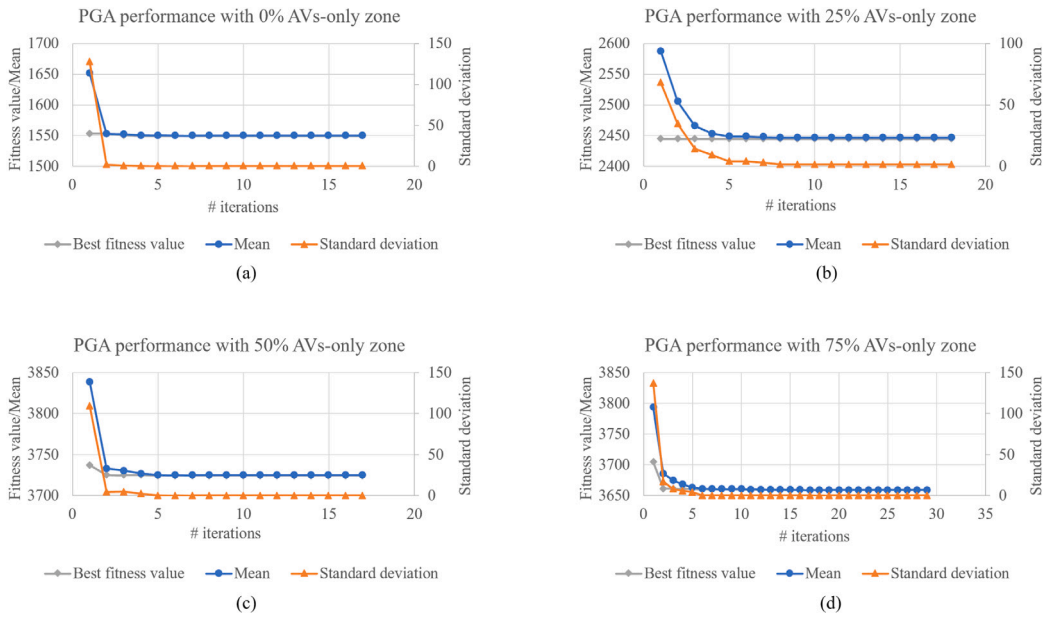

Fig. 7. Performance of the solution method with different coverage rates of the AVs-only zone: best fitness value, mean and standard deviation of the fitness value of the top five fittest individuals.

5.2.3. Comparison between the best fleet size and the lower bound

The optimisation results for the base scenario are shown in Table 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 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 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 paper.

5.2.4. 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 Fig. 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%, and 75%) and the best fleet sizes.

As illustrated in Fig. 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.

5.2.5. 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 ± 3 time steps (a total time range of 15 min) 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

Table 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		15 539.92	15 601.70	16 803.63	16 861.36	17 643.82	17 674.55
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

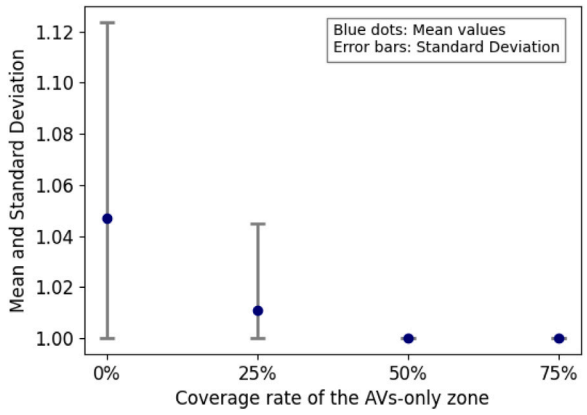

Fig. 8. Mean and standard deviation of the cost ratios across all groups of trips for scenarios with different coverage rates of AVs-only zone (0%, 25%, 50%, and 75%) and the best fleet sizes.

with different datasets is displayed in Fig. 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.

The computational results are shown in Table 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 towards 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.

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 Section 5.2.3 that the smallest possible fleet size of CTs is always preferable for a TNC in this study.

5.2.6. *Impacts of AVs-only zones*

The upgrade of the conventional road networks to AVs-only zones brings inevitable effects on the demand patterns, ride-hailing operations, behaviours of travellers, and traffic conditions on road networks. Table 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 11 show a significant increase in the relocation distance share of CTs' total travel distance when the

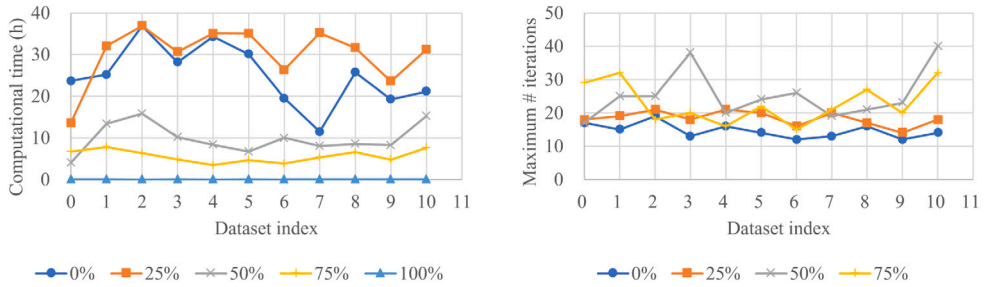

Fig. 9. Performance of the solution method with different datasets: (a) Computational time, (b) Maximum number of iterations.

Table 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

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 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 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 is 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.

6. Main conclusions and future work

Envisioning the emergence and expansion of AVs-only zones in urban areas, a bi-level framework has been proposed in this paper 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 is 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Chinese Scholarship Council (CSC).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.01.017.

References

Allahviranloo, M., & Chow, J. Y. (2019). A fractionally owned autonomous vehicle fleet sizing problem with time slot demand substitution effects. *Transportation Research Part C (Emerging Technologies)*, *98*, 37–53.

Ashkrof, P., Homem Correia, G., Cats, O., & Van Arem, B. (2022). Ride acceptance behaviour of ride-sourcing drivers. *Transportation Research Part C (Emerging Technologies)*, *142*, Article 103783.

Bagloee, S. A., Sarvi, M., Patriksson, M., & Rajabifard, A. (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), 562–580.

Balac, M., Hörl, S., & Axhausen, K. W. (2020). Fleet sizing for pooled (automated) vehicle fleets. *Transportation Research Record*, *2674*(9), 168–176.

Bösch, P. M., Becker, F., Becker, H., & Axhausen, K. W. (2018). Cost-based analysis of autonomous mobility services. *Transport Policy*, *64*, 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), 716–728.

Camacho-Vallejo, J.-F., López-Vera, L., Smith, A. E., & González-Velarde, J.-L. (2021). A tabu search algorithm to solve a Green logistics bi-objective bi-level problem. *Annals of Operations Research*, 1–27.

Chen, Z., He, F., Yin, Y., & Du, Y. (2017). Optimal design of autonomous vehicle zones in transportation networks. *Transportation Research, Part B (Methodological)*, *99*, 44–61.

Chen, Z., He, F., Zhang, L., & Yin, Y. (2016). Optimal deployment of autonomous vehicle lanes with endogenous market penetration. *Transportation Research Part C (Emerging Technologies)*, *72*, 143–156.

Chen, R., & Levin, M. W. (2019). Dynamic user equilibrium of mobility-on-demand system with linear programming rebalancing strategy. *Transportation Research Record*, *2673*(1), 447–459.

Chen, A., Pravinvongvuth, S., Xu, X., Ryu, S., & Chootinan, P. (2012). Examining the scaling effect and overlapping problem in logit-based stochastic user equilibrium models. *Transportation Research Part A: Policy and Practice*, *46*(8), 1343–1358.

Chondrogiannis, T., Bouros, P., Gamper, J., Leser, U., & Blumenthal, D. B. (2020). Finding k-shortest paths with limited overlap. *The VLDB Journal*, *29*(5), 1023–1047.

Conceição, L., Correia, G., & Tavares, J. P. (2021). Automated vehicles (AV) dedicated networks and their effects on the traveling of conventional vehicle drivers. *Transportation Research Procedia*, *52*, 653–660.

Correia, G. H., Looff, E., Van Cranenburgh, S., Snelder, M., & Van Arem, B. (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*, 359–382.

Correia, G. H., & Van Arem, B. (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*, 64–88.

Dafermos, S. C., & Sparrow, F. T. (1969). The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards B*, *73*(2), 91–118.

Eklund, S. E. (2004). A massively parallel architecture for distributed genetic algorithms. *Parallel Computing*, *30*(5–6), 647–676.

Fagnant, D. J., & Kockelman, K. M. (2018). Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation*, *45*(1), 143–158.

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

Farahani, R. Z., Miandoabchi, E., Szeto, W. Y., & Rashidi, H. (2013). A review of urban transportation network design problems. *European Journal of Operational Research*, *229*(2), 281–302.

Ge, Q., Han, K., & Liu, X. (2021). Matching and routing for shared autonomous vehicles in congestible network. *Transportation Research Part E: Logistics and Transportation Review*, *156*, Article 102513.

Guo, Q., Ban, X. J., & Aziz, H. A. (2021). Mixed traffic flow of human driven vehicles and automated vehicles on dynamic transportation networks. *Transportation Research Part C: Emerging Technologies*, *128*, Article 103159.

Guo, Z., Hao, M., Yu, B., & Yao, B. (2021). 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*, Article 103390.

Hiermann, G., Puchinger, J., Ropke, S., & Hartl, R. F. (2016). The electric fleet size and mix vehicle routing problem with time windows and recharging stations. *European Journal of Operational Research*, *252*(3), 995–1018.

Hoang, N. H., Panda, M., Vu, H. L., Ngoduy, D., & Lo, H. K. (2023). A new framework for mixed-user dynamic traffic assignment considering delay and accessibility to information. *Transportation Research Part C (Emerging Technologies)*, *146*, Article 103977.

Kashmiri, F. A., & Lo, H. K. (2022). Routing of autonomous vehicles for system optimal flows and average travel time equilibrium over time. *Transportation Research Part C (Emerging Technologies)*, *143*, Article 103818.

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, *80*(5), 8091–8126.

Ke, Z., & Qian, S. (2023). Leveraging ride-hailing services for social good: Fleet optimal routing and system optimal pricing. *Transportation Research Part C (Emerging Technologies)*, *155*, Article 104284.

Koç, Ç., Bektaş, T., Jabali, O., & Laporte, G. (2016). The fleet size and mix location-routing problem with time windows: Formulations and a heuristic algorithm. *European Journal of Operational Research*, *248*(1), 33–51.

Kouwenhoven, M., de Jong, G. C., Koster, P., Van den Berg, V. A., Verhoef, E. T., Bates, J., & Warffemius, P. M. (2014). New values of time and reliability in passenger transport in the netherlands. *Research in Transportation Economics*, *47*, 37–49.

Laporte, G. (2009). Fifty years of vehicle routing. *Transportation Science*, *43*(4), 408–416.

Li, Q., & Liao, F. (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*, 151–175.

Li, R., Liu, X., & Nie, Y. M. (2018). Managing partially automated network traffic flow: Efficiency vs. stability. *Transportation Research, Part B (Methodological)*, *114*, 300–324.

Liang, X., Correia, G. H., An, K., & Van Arem, B. (2020). Automated taxis' dial-a-ride problem with ride-sharing considering congestion-based dynamic travel times. *Transportation Research Part C (Emerging Technologies)*, *112*, 260–281.

Liang, X., Homem Correia, G., & Van Arem, B. (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), 588–598.

Liang, Q., Li, X.-a., Chen, Z., Pan, T., & Zhong, R. (2023). Day-to-day traffic control for networks mixed with regular human-piloted and connected autonomous vehicles. *Transportation Research, Part B (Methodological)*, *178*, Article 102847.

Liu, S., Huang, W., & Ma, H. (2009). An effective genetic algorithm for the fleet size and mix vehicle routing problems. *Transportation Research Part E: Logistics and Transportation Review*, *45*(3), 434–445.

Liu, H., Jin, C., Yang, B., & Zhou, A. (2017). Finding top-k shortest paths with diversity. *IEEE Transactions on Knowledge and Data Engineering*, *30*(3), 488–502.

Liu, J., Mirchandani, P., & Zhou, X. (2020). Integrated vehicle assignment and routing for system-optimal shared mobility planning with endogenous road congestion. *Transportation Research Part C (Emerging Technologies)*, *117*, Article 102675.

Liu, Z., & Song, Z. (2019). Strategic planning of dedicated autonomous vehicle lanes and autonomous vehicle/toll lanes in transportation networks. *Transportation Research Part C (Emerging Technologies)*, *106*, 381–403.

Madadi, B., Van Nes, R., Snelder, M., & Van Arem, B. (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), 80–96.

Mansourianfar, M. H., Gu, Z., & Saberi, M. (2022). Distance-based time-dependent optimal ratio control scheme (TORCS) in congested mixed autonomy networks. *Transportation Research Part C (Emerging Technologies)*, *141*, Article 103760.

Mansourianfar, M. H., Gu, Z., Waller, S. T., & Saberi, M. (2021). Joint routing and pricing control in congested mixed autonomy networks. *Transportation Research Part C (Emerging Technologies)*, *131*, Article 103338.

Militão, A. M., & Tirachini, A. (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*, 52–80.

Mo, D., Chen, X. M., & Zhang, J. (2022). Modeling and managing mixed on-demand ride services of human-driven vehicles and autonomous vehicles. *Transportation Research, Part B (Methodological)*, *157*, 80–119.

Nieuwenhuijsen, M. J., & Khreis, H. (2016). Car free cities: Pathway to healthy urban living. *Environment International*, *94*, 251–262.

Olia, A., Razavi, S., Abdulhai, B., & Abdelgawad, H. (2018). Traffic capacity implications of automated vehicles mixed with regular vehicles. *Journal of Intelligent Transportation Systems*, *22*(3), 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.

Renaud, J., & Boctor, F. F. (2002). A sweep-based algorithm for the fleet size and mix vehicle routing problem. *European Journal of Operational Research*, *140*(3), 618–628.

Santos, G. G. D., & Correia, G. M. (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*, Article 103092.

Scherr, Y. O., Saavedra, B. A. N., Hewitt, M., & Mattfeld, D. C. (2019). Service network design with mixed autonomous fleets. *Transportation Research Part E: Logistics and Transportation Review*, *124*, 40–55.

Schnabel, W., & Löhse, D. (1997). *Grundlagen der Strasssen-Verkehrstechnik und der Verkehrsplanung, Band 2, neue bearb*. Aufl.

Sheffi, Y. (1985). *Urban transportation networks, vol. 6*. Englewood Cliffs, NJ: Prentice-Hall.

Stoiber, T., Schubert, I., Hoerler, R., & Burger, P. (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*, 265–282.

UITP (2017). Autonomous vehicles: a potential game changer for urban mobility. In *Policy brief*. Brussels: International Association of Public Transport (UITP).

Van Essen, J. T., & Correia, G. H. (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), 4567–4578.

Vega-Gonzalo, M., Aguilera-García, Á., Gomez, J., & Vassallo, J. M. (2023). Traditional taxi, e-hailing or ride-hailing? a gsem approach to exploring service adoption patterns. *Transportation*, 1–40.

Wang, S., Correia, G. H. d. A., & Lin, H. X. (2022). Assessing the potential of the strategic formation of urban platoons for shared automated vehicle fleets. *Journal of Advanced Transportation*.

Wang, Z., Qi, M., Cheng, C., & Zhang, C. (2019). A hybrid algorithm for large-scale service network design considering a heterogeneous fleet. *European Journal of Operational Research*, *276*(2), 483–494.

Yang, K., Guler, S. I., & Menendez, M. (2016). Isolated intersection control for various levels of vehicle technology: Conventional, connected, and automated vehicles. *Transportation Research Part C (Emerging Technologies)*, *72*, 109–129.

Yang, K., Tsao, M. W., Xu, X., & Pavone, M. (2020). Planning and operations of mixed fleets in mobility-on-demand systems. arXiv preprint arXiv:2008.08131.

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), 526–530.

Yi, Z., & Smart, J. (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*, Article 102822.

Zhang, F., Lu, J., & Hu, X. (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*, Article 102906.

Zhang, K., & Nie, Y. M. (2018). Mitigating the impact of selfish routing: An optimal-ratio control scheme (ORCS) inspired by autonomous driving. *Transportation Research Part C (Emerging Technologies)*, *87*, 75–90.