Economics of Transportation 34 (2023) 100313

Contents lists available at ScienceDirect

Economics of Transportation

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

Economies and diseconomies of scale in on-demand ridepooling systems

Andrés Fielbaum [a,b,*], Alejandro Tirachini [c,d], Javier Alonso-Mora [b]

[a] School of Civil Engineering, University of Sydney, Australia
[b] Department of Cognitive Robotics, TU Delft, Netherlands
[c] Department of Civil Engineering and Management, University of Twente, Enschede, the Netherlands
[d] Civil Engineering Department, Universidad de Chile, Santiago, Chile

ARTICLE INFO

Keywords:
On-demand mobility
Ridepooling
Scale economies
Mohring effect
Extra-detour effect
Better-matching effect
Automated vehicles

ABSTRACT

We analyse the sources of economies and diseconomies of scale in On-Demand Ridepooling (ODRP), disentangling three effects: when demand grows, average costs are reduced due to *i*) a larger fleet that diminishes waiting and walking times (*Mohring Effect*), and *ii*) matching users with more similar routes (*Better-matching Effect*). A counter-balance force (*Extra-detour Effect*), occurs when *iii*) the number of passengers per vehicle increases and users face longer detours. At low demand levels, there is little sharing and the Mohring effect prevails; as demand grows, more passengers per vehicle push for the Extra-detour Effect to dominate; eventually, vehicles run at capacity, and the Better-matching Effect prevails. The last two effects are specific to ODRP as the routes are not fixed but adapted online. Our simulations show that considering both users' and operators' costs, scale economies prevail, and that ODRP with human-driven vehicles and walks allowed has total costs similar to door-to-door systems with driverless vehicles.

1. Introduction

1.1. On-demand ridepooling systems: potential and challenges

Transport systems are facing profound transformations worldwide thanks to the ability to connect vehicles and large numbers of passengers on demand. After almost ten years since their arrival, several studies have shown that transportation network companies (TNCs) based on unshared rides (also called ride-hailing or ridesourcing) have increased traffic and congestion (Diao et al., 2021; Henao and Marshall, 2019; Roy et al., 2020; Tirachini and Gomez-Lobo, 2020; Ward et al., 2021; Wu and MacKenzie, 2021). This situation has fostered the study and implementation of on-demand *ridepooling* (ODRP) services, in which different users simultaneously share a vehicle when their routes are compatible, so that congestion and emissions might be reduced (Li et al., 2021; Tikoudis et al., 2021), depending on modal substitution.

ODRP systems have the potential to lower congestion because they might reduce the required fleet significantly when compared to the non-pooled versions, as shown by several previous studies (Alonso-Mora et al., 2017; Fagnant & Kockelman, 2018; Santi et al., 2014). However, these analyses are based on comparing the number of vehicles needed to serve a fixed demand level, which might be a strong assumption as both types of service do not necessarily attract the same number of users. In fact, recent studies suggest that the ability of ODRP to reduce congestion depends on reaching some advantageous scenarios (Ke et al., 2020; Tirachini et al., 2020). Such scenarios should combine an efficient fleet operation with an ability to attract a significant number of passengers from private cars. To reach those scenarios, some strategic decisions arise, such as whether it is efficient to use ODRP in low-demand or high-demand markets, or if it should replace and/or complement a public transport service over a network.

These strategic questions require a deeper understanding of the user and operator cost structure of ODRP systems, in particular, of the mechanisms that introduce economies or diseconomies of scale. However, this is not an easy task, as the operation of ODRP depends on specific algorithms to face the complexity of operating on demand and with a large number of feasible ways to match users and vehicles. Which algorithm to utilise may yield different strategic results and affect scale effects. For instance, a seminal study by Li and Quadrifoglio (2010) studies a last-mile service that dispatches vehicles sequentially as soon as they get enough users regardless of their destinations. When doing so, a potential source of scale economies is not leveraged, namely that a greater demand enables grouping together users with closer destinations without increasing waiting times significantly. This is the type of issue

* Corresponding author. School of Civil Engineering, University of Sydney, Australia
E-mail address: andres.fielbaum@sydney.edu.au (A. Fielbaum).

https://doi.org/10.1016/j.ecotra.2023.100313
Received 11 August 2022; Received in revised form 14 April 2023; Accepted 26 April 2023
Available online 13 May 2023

that we address in this paper.

1.2. Overview, contributions, and structure of the paper

In this paper, we extend a state-of-the-art assignment model to perform a detailed economic analysis of ODRP systems. In particular, we uncover three sources of economies and diseconomies of scale that are present simultaneously when operating ODRP, with the objective to determine the efficiency of ODRP as a shared-mobility platform for urban operations.

To do so, we consider a real-life network in Manhattan, New York, and also an extended version of the so-called single-line model (borrowed from public transport), in order to study sources of economies and diseconomies of scale when operating ODRP. The traditional single-line model has been extensively used by researchers across decades to analyse structural aspects of mass public transport design. The model is useful to study the impact of the demand levels, values of time, operator costs and other parameters over the mobility system under scrutiny, simplifying its spatial distribution. By this means, the demand can be represented by a single variable (or a few variables), usually in passengers per hour, which makes this model quite precise for scale analysis (e.g., Fielbaum et al., 2020a; Tirachini and Antoniou, 2020).

The single-line model is useful for scale analysis, but has a relevant limitation when studying on-demand systems: the on-demand nature entails that vehicles' routes should not be defined a priori but adapted to the emerging users. Such a feature cannot be captured in a single-line model in which there is only one possible route. This limitation might influence scale analysis, as one aspect to study is the evolution of the routes with scale (in fact, Manik and Molkenthin, 2020, show that a linear network artificially favours the performance of ODRP over several alternative topologies). We address this limitation by including simulations using a real-life dataset from Manhattan, and by extending the single-line model, so that we keep most of its simplifying aspects, but yet enabling different routes to be followed depending on the passengers. In simple terms, we deploy a grid surrounding the single-line, so that the vehicles move within the grid depending on the specific users they are serving.

In our setting, we have another challenge that arises when analysing scale for on-demand systems: which fleet to use. Most models that simulate ODRP assume a given fleet (as we describe further in Section 2.1). However, a proper scale analysis requires that the fleet is endogenously computed, which is troublesome in ODRP as even with a fixed fleet the mathematical problems present great complexity. Here we propose a method to compute the fleet together with the assignment decisions, so that the number of vehicles responds to the demand.

We are interested in the potential of ODRP to face some of the most relevant challenges faced by urban transport, such as emissions and congestion externalities, so the system we study follows rules that resemble public transport operations. It is non-profit, and the costs of all agents (users and operators) are considered when deciding how to assign vehicles to users. The number of vehicles and their routes are decided by a central controller aiming to minimise a function that represents total costs, where we impose that all users must be served. Moreover, we do not impose a door-to-door service, i.e., the system might decide (on-demand) pick-up and drop-off points that require some short walks, if doing so improves the system's overall performance. That walking time has a valuation for the user that is different from the valuation of in-vehicle time.

Our main contribution is to theoretically disentangle and discuss in depth three sources of economies and diseconomies of scale in ODRP systems, which are then verified through numerical simulations under several different scenarios. These scale effects illustrate the potential and obstacles that need to be overcome for ODRP to succeed (for instance, Bahrami et al., 2022 show that the profitability of ODRP depends on the presence of scale economies when matching different users). Some of these sources of economies and diseconomies of scale are specific to ODRP systems, as they depend on how the flexible routes followed by the vehicles evolve when the number of passengers grows. Furthermore, we propose a way to compute the fleet size in ODRP together with the assignment decisions, which can be utilised for other types of analysis beyond the objectives of this paper. We also show the potential of relaxing the door-to-door scheme when all requests must be served, and compare our results with an idealised public transport system.

As discussed exhaustively in Section 3, there are studies that identify some of the scale effects of ODRP analysed in the present research (Daganzo et al., 2020; Kaddoura and Schlenther, 2021; Ke et al., 2020; Lehe et al., 2021, Militão and Tirachini, 2021, Zhang and Nie, 2021), although most of these works focus on a single effect. Compared to this body of research, our first contribution is the setting of a single framework that allows us to identify and combine the three sources of economies and diseconomies of scale previously mentioned, in an integrated fashion that reveals which of them dominates as the demand grows. Second, in our model the fleet is endogenously adjusted to the increasing demand levels, which can be compared to previous research efforts that usually assumed a fleet size exogenously given. Third, in the application of the model, we compare alternative deployment scenarios in order to address important policy and service design questions, such as what the efficiency gains of allowing walks to pick-up and from drop-off points are, compared to door-to-door operation systems (for both human-driven and driverless vehicles), and what are the implications of alternative operation rules in the cost comparison between ODRP and fixed-route public transport.

The paper is organised as follows. Section 2 revises relevant previous studies. Section 3 describes qualitatively and formalises which are the most relevant novel sources of scale economies and diseconomies that emerge for a transport system that is both on-demand and shared. Section 4 shows the numerical simulations, for which we first explain the methodological challenges and how we face them. Finally, Section 5 concludes and proposes some directions for future research.

2. Related works

2.1. Fleet sizing in on-demand ridepooling systems

Deciding the fleet size to be used in an ODRP system is not an easy task. Contrary to public transport, the routes cannot be known in advance, so the usual techniques dealing with cycle times and desired frequencies (see Jara-Díaz and Gschwender, 2003 for a survey on this topic) cannot be applied here. Furthermore, the ideas that have been used for non-shared on-demand systems, where the crucial question is how to chain consecutive trips (such as Vazifeh et al., 2018), are also not applicable in this context, because here the trips of different users overlap. Such difficulties have been faced with different approaches that we now describe.

The most usual approach in the operations research literature is to work with fleets of fixed size. In order to determine which fleet size is optimal, or at least gain some intuition about this issue, it is common to repeat the same numerical experiments with different fleet sizes, to analyse which size adjusts more efficiently to a given demand level, by comparing some metrics on, e.g., operating costs, share of unserved demand, waiting time and travel time (Alonso-Mora et al., 2017; Levin et al., 2017; Lokhandwala and Cai, 2018; Wang et al., 2018). Other studies seek the minimal fleet able to meet some exogenous conditions on the quality of service. For instance, Daganzo and Ouyang (2019) and Martinez and Viegas (2017) require to serve all the demand, although the latter also compare to the results obtained with larger fleets. Spieser et al. (2014) consider bounds on the number of passengers waiting to be served, and Fagnant & Kockelman (2018) aim at fulfilling some predefined waiting times.

Alternative rules to analyse fleet size in ODRP include the proposals of Santos and Xavier (2015), who assume that the number of vehicles

has to be proportional to the number of requests, a rule that is obtained as a result by Kang and Levin (2021) when following an assignment policy that aims at maximising the number of users per vehicle; Pinto et al. (2020), who consider an available budget, shared with public transport, that has to be respected; and Fielbaum (2020), who makes a weighted optimization between users' and operators' costs under simplifying assumptions that lead to the prediction of exact fleet sizes. Cáp and Alonso-Mora (2018) explain that the optimal fleet size also considers both types of costs and study the corresponding multi-objective problem, proposing a method to compute the Pareto front.

2.2. The single-line model and other simplified networks for the analysis of transport systems

The single line model has been used to identify scale effects in public transport. The stream of studies based on the single-line model was pioneered by Mohring (1972), who identified one of the main sources of scale economies in public transport (now known as the "Mohring Effect"): more passengers require more buses, which increases the service frequency and diminishes waiting times for everybody. His model was later extended by Jansson (1980) to consider optimal bus capacities and time at stops, where a source of diseconomies of scale emerges, namely that an increase in the number of users yields the utilisation of larger buses, making users to spend more time waiting for other passengers to board and alight (an effect that can be compensated by changing the number of doors per vehicle, as analysed by Jara-Díaz and Tirachini, 2013). Evans and Morrison (1997) discovered yet another source of scale economies with an extension of this model: an increase in the number of users enables spending more resources in preventing accidents and disruptions in the service. Finally, the inclusion of a crowding externality as increasing the value of in-vehicle time savings for public transport users has been shown to increase average costs (and therefore introduce diseconomies of scale) for large demand levels in a single-route model (Tirachini et al., 2010a), however, when the number of routes can be optimally decided to minimise total costs, route density is increased to reduce crowding and keep total costs down even for large demand levels (Tirachini et al., 2010b). Most of these effects have been shown to remain valid for single lines when a network is considered (Fielbaum et al., 2020a).

Similar analyses have also been conducted in other networks, although finding a suitable simplified representation is already a complex task (see Fielbaum et al., 2017, for an exhaustive discussion on this issue). In the context of ODRP, Pimenta et al. (2017) has utilised a single-line model to discuss how to operate the system in a reliable way; Badia and Jenelius (2021) and Chen and Nie, (2017) consider simplified grids to study how to connect ODRP with mass public transport, a problem that is studied over an homogeneous circle by Fielbaum (2020); while Manik and Molkenthin (2020) compare different simplified topologies to analyse which of them are better served through ODRP.

2.3. Comparing ridepooling and public transport

Previous studies have compared whether it is more efficient to utilise ODRP instead of fixed-route public transport services in a given area; the usual result is that ODRP is more convenient only for low-demand services. , as well as Papanikolaou and Basbas (2020), have rested on specific functional forms that approximate the operation of ODRP systems, finding that ODRP should be preferred not only when the demand is low but also when the areas to be served are small and trips are short. Quadrifoglio and Li (2009) and Li and Quadrifoglio (2010) use continuous approximation models and identify the discomfort of walking as another relevant parameter that determines which type of system should be preferred. Similarly, Calabrò et al. (2021) use microsimulation to find that flexible services are better in rural areas. On the contrary, Bischoff et al. (2019) suggest that public transport could be fully replaced by ODRP in small or medium cities, while Viergutz and Schmidt (2019) conclude that rural areas should use line-based on-demand services rather than completely flexible routes.

It should be noted that all these models assume that the flexible systems provide door-to-door service (or station-to-door, when it is solving the last-mile problem), which is a common assumption as most real-life on-demand systems operate in that way. However, operating door-to-door is not mandatory for this type of system. Actually, previous research has consistently shown that requesting some users to walk either to personalised pick-up and drop-off points (Fielbaum, 2021; Fielbaum et al., 2021; Lotze et al., 2022; Wang et al., 2022) or to group meeting points (Bischoff et al., 2019; Li et al., 2016, 2018; Stiglic et al., 2015) can enhance ODRP services significantly. Such ideas are already applied in real life: for instance the shared-mobility platform Jetty in Mexico City asks passengers to be at specific pick-up points to be able to board a shared car or van; and users can monitor the location of the vehicle in real-time before boarding (Tirachini et al., 2020).

3. Sources of scale economies in ODRP

3.1. Definition of an ODRP cost function

The problem of operating an ODRP system is defined by an urban environment (usually a network represented by a directed graph), a fleet of vehicles V, and a demand consisting of a set of requests $r \in R$. Each request is characterised by its origin, destination and the time in which the request was placed. Crucially, the demand is not known beforehand; instead, the system can only take a request into account when it emerges, i.e., the time in which the request was placed. A solution to this problem is defined by a route Π_v for each vehicle v, that serves a set of requests R_v, in such a way that the capacity of the vehicle is never exceeded, and without violating other constraints that could be defined by the operator or service manager (such as hard restrictions on total waiting and travel times).

This general problem can be formulated in many different ways. In computational complexity theory, the ODRP problem combines two well-known NP-Hard problems, namely Dynamical-Vehicle-Routing-Problem and Dial-A-Ride (Yu and Shen, 2020). Therefore, in the past years, several methods and heuristics have been proposed to operate ODRP systems and determine how to decide routes and assign trip requests to vehicles. When solving the problem, researchers usually follow a batch-based approach, in which emerging requests are accumulated during some lapse of time before deciding how to assign them all together (e.g., Alonso-Mora et al., 2017; Simonetto et al., 2019; Tsao et al., 2019), or an event-based approach, where each request is assigned as soon as it appears (Fagnant & Kockelman, 2018; Militão and Tirachini, 2021; van Engelen et al., 2018). When deciding routes and assignments, all of these techniques consider some objective function, i.e., there is an implicit or explicit *cost function* that the model tries to minimise.

What is the economic meaning of the cost function in the transport context? As discussed by Jara-Diaz (2007), the *product* in transport systems is defined by the demand being transported, and the problem is how to serve it optimally, i.e., how to define the fleet composition, routes, and assignments, to minimise a certain cost function. Which elements should be accounted for when defining the ODRP cost function? First, we should note that the standard production approach to scale economies that is found in other markets (i.e., how do average production costs evolve when exogenous output increases) is incomplete to analyse passenger transport systems as ODRP. In the public transport literature, such realisation came in the 70s, with the pioneering works of Mohring (1972), Turvey and Mohring (1975), and Jansson (1979), who were among the first to argue that all users' time costs and efforts should be treated as costs on a par with operators' costs, when analysing the optimal design and pricing of public transport systems, because considering operators' costs only leads to suboptimal fleets (e.g., too few

vehicles that increase waiting time, too small vehicles that increase passengers' crowding). The relevance of including both users' costs and operators' costs in the analysis of public transport provision has been subsequently exposed by many studies (see reviews by Jara-Díaz and Gschwender, 2003; Hörcher and Tirachini, 2021). In what follows, we adopt this paradigm of total cost functions - including users and operators-for the analysis of ODRP systems.

Operator costs are defined by capital and operating costs (Delle Site and Filippi, 1998; Jara-Díaz et al., 2017), which depend namely on the fleet composition (number of vehicles B [veh] and their size K [pax/veh]), and their usage (defined by the vehicles-hour-travelled VHT or the vehicles-kilometres-travelled VKT), respectively. User costs are more difficult to define, as they aim to capture all the subjective aspects of users' experience. User costs should at least consider the average times involved in the different stages of the transport process: waiting time t_w [min], walking time t_a [min], and in-vehicle time t_v [min]. Other aspects that can be included, but are disregarded in this paper, are the unreliability of the system, how comfortable it is, or the eventual (and undesirable) presence of transfers.[1] Putting everything together, the cost function can be written as:

$$cost = c_O(B, K) + c_U(t_w, t_a, t_v) \tag{1}$$

Where c_O and c_U stand for operators' and users' costs, respectively. A usual approach (similar to what we do when running simulations in Section 4) is to assume these functions as $c_O = (c_{O1} + c_{O2}K)B$ and $c_U = p_w t_w + p_a t_a + p_v t_v$, where c_{O1}, c_{O2} are operator cost parameters that translate everything into the same monetary currency, and p_w, p_a and p_v are the value of waiting, access and in-vehicle time, respectively. The resulting users and operators costs are not exogenous, as they are endogenously obtained when minimising the cost function given by Eq. (1). For instance, when deciding which vehicle to assign to a particular request, operating costs as well as waiting and in-vehicle time costs need to be considered. Investigating scale economies in these systems refers precisely to all the sources of costs in Eq. (1). Overall scale economies means that the average total costs decrease as the number of users increases, and this analysis can be disentangled to analyse what happens with each component of the cost function when the demand grows.

As discussed by Basso and Jara-Díaz (2006), scale analysis in transport systems is a complex task because the demand has a spatial dimension that cannot be aggregated through simplified indices such as Passenger-Kilometres. Our analysis should be interpreted as a *ray analysis* (Baumol, 1986), i.e., the demand grows proportionally keeping the spatial distribution constant. This implies that the demand (i.e., the *product*) can be described by a single variable Y [pax/hour]. When we run numerical simulations (Section 4), such a ray analysis is achieved first through the utilisation of (an expanded version of) the single-line model, and then via selecting an increasing number of random requests from a real-life dataset in Manhattan. It is worth noting that when demand grows, its distribution might actually change, so that the scale analysis done here should be complemented with the analysis of economies of scope (Jara-Diaz, 2007). Moreover, the on-demand nature of ODRP increases the difficulty of the study of scale economies, as the traditional techniques presume that the decisions (here the fleet size and their routes) are taken optimally, but doing so might be unfeasible when the demand is not known beforehand, which is why we rely on a thorough set of simulations in Section 4 to verify and compare the different scale effects introduced later in this section.

We now propose and explain three sources of scale economies dealing with users, which will be studied numerically in Section 4. Let us denote by Y the number of users per hour in the system, and by ρ [pax/veh] the average load -or occupancy rate-of the vehicles (note that, by definition, it must hold that $\rho \leq K$), which refers to the average number of passengers on board of a vehicle. As the number of vehicles and their operation is decided optimally,[2] the variables B and ρ should respond endogenously to the demand, i.e., $B = B(Y), \rho = \rho(Y)$. The fleet size and the occupancy of vehicles critically influence users' experience and satisfaction, meaning that $t_h(Y) = t_h(B(Y), \rho(Y), Y)$ for $h = w, a, v$. We now study the effect of each of these variables on waiting, access, and in-vehicle times.

3.2. The extra-detour effect

When more users enter an ODRP system, it becomes usually possible to serve more of them simultaneously with the same vehicle, even if there are restrictions on the total waiting or travel times.[3] Pooling users in shared rides is often optimal, as fewer vehicles are required compared to an alternative with less sharing. Formally speaking, this means that $\rho'(Y) > 0$. **The Extra-detour effect is a source of scale diseconomies for users, defined as the degradation in the quality of service due the extra detours induced by the increase in the average number of passengers per vehicle.** Intuitively, as the vehicle routes are not defined a priori but adapted to the specific users being served, the quality of service perceived by the users is sensitive to route choice. When the vehicles are more shared, this increases the detours required by the system (Militão and Tirachini, 2021a), which in turn increases waiting times. Moreover, the chance of walking instead of having a door-to-door service increases as well, because the time savings from walking are larger when more other passengers are affected. The Extra-detour effect can be expressed mathematically as[4]:

$$\frac{\partial t_h}{\partial \rho} \text{ for } h = w, a, v \tag{2}$$

The Extra-detour effect is illustrated in Fig. 1, where we show how the blue passenger increases all the components of her travel time when the vehicle serves a new user (red). The Extra-detour effect can get exhausted when vehicles run at capacity (or near capacity).

As discussed by Fielbaum and Alonso-Mora (2020), the fact that routes are not known beforehand, but depend on the travellers, is specific to mobility providers that are both shared (otherwise vehicles follow shortest paths) and on-demand (otherwise vehicles follow fixed routes). Therefore, this source of scale diseconomies is specific to ODRP systems. Nevertheless, the Extra-detour effect can be interpreted as similar to a well-known fact in public transport, namely that new users increase the vehicle occupancy rate, which in turn increases the time spent at stops waiting for boarding and alighting passengers.

It is noteworthy that the Extra-detour effect can affect other service attributes, besides t_w, t_a and t_v:

- Fielbaum and Alonso-Mora (2020) identify two types of **unreliability** in ODRP**:** The "one-time unreliability", defined as changes that take place while a trip is executed due to emerging requests, and

[1] For the case of public transport, van Lierop et al. (2018) provide a review on the factors considered by users when evaluating their experience.

[2] Or following some heuristic aiming for optimal decisions. We note that for this analysis, we assume K to be exogenous, and we show in Section 4 that the analysis remains valid if K was also optimised.

[3] This can be formalised through the so-called *shareability networks*, that measure how many requests can be combined together. As shown by Santi et al. (2014), and Tachet et al. (2017), a greater number of travellers entails a larger shareability.

[4] Note that there might be specific circumstances in which a vehicle's load can increase without increasing service times (for instance, starting from any base situation, and duplicating the number of users for all those requests that have low waiting times). Eqs. (2)–(4) are valid when the demand grows without changing the spatial distribution. For a thorough discussion about the methodological challenges of studying scale economies in transport systems when the network and/or spatial distribution of the demand can be changed, see Basso and Jara-Díaz (2006).

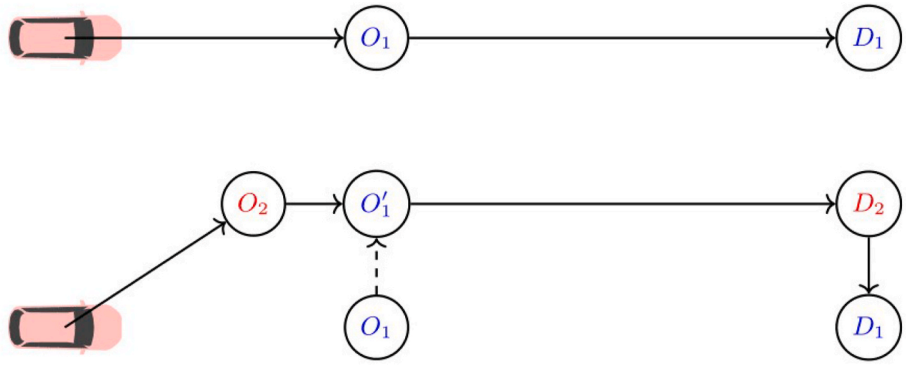

Fig. 1. Example of the Extra-detour effect. The number of passengers is low in the top row, so users do not share the vehicle, and the blue passenger faces little waiting time and no detour. When the demand grows (bottom row), a new red co-traveller appears close to her, which increases her waiting time, requires her to walk (marked with a dotted arrow), and implies a detour, degrading her perceived quality of service.

the "daily unreliability", that refers to facing different conditions each time a trip is repeated. Both types of unreliability worsen when vehicles are more loaded, i.e., the Extra-detour Effect increases unreliability as well. This is not a minor issue: for instance, Alonso-González et al. (2020) have estimated the value of reliability (that refers to the daily unreliability discussed above) to be approximately half of the value of time.

- Sharing the vehicle with more users can be uncomfortable by itself, as studied by Ho et al. (2018), König and Grippenkoven (2020), and Lavieri and Bhat (2019), who propose the so-called "**willingness to share**" to study the difference in comfort between travelling alone or with other users. Note that this effect only occurs when vehicles start to increase their number of passengers. The willingness or unwillingness to share a ride is related to an increase in **crowding**, i.e., the discomfort perceived by passengers when having to share a limited space (a vehicle or a station) with a large amount of passengers, which has been thoroughly studied in the public transport literature, as surveyed by Tirachini et al. (2013).

Although the Extra-detour effect is undesirable for passengers, there is one positive consequence of the fact that $\rho'(Y) > 0$ from the operators' standpoint: namely, the number of vehicles per hundred users diminishes thanks to an increase in vehicle usage. This source of scale economies is usual in shared systems (Fielbaum et al., 2020a).

3.3. The Better-matching effect

Some papers that analyse ODRP have reported that, when the demand is large enough, it becomes possible (and thus optimal) to form more efficient groups of users (Daganzo et al., 2020; Ke et al., 2020; Lehe et al., 2021, Zhang and Nie, 2021). This is an intuitive result, as a larger demand implies that there are more feasible requests that can be matched together. The Better-matching effect is thus defined as **the ability to create groups whose routes are more compatible with each other when the number of passengers increases, thanks to a larger pool of requests to choose from.** Formally:

$$\frac{\partial t_h}{\partial Y} \leq 0 \text{ for } h = w, a, v \tag{3}$$

The Better-matching effect is illustrated in Fig. 2, where users 1 and 2 are first grouped together; when new passengers emerge, they are separated and matched with other users such that the resulting routes get more efficient.

The Better-matching effect also emerges thanks to the flexibility of the routes, i.e., it is specific to ODRP systems. However, it is similar to the increase in "directness" in public transport systems reported by Fielbaum et al. (2020a), who argue that an increased number of passengers enables the definition of lines that require fewer detours because more passengers share the same origins and destinations.

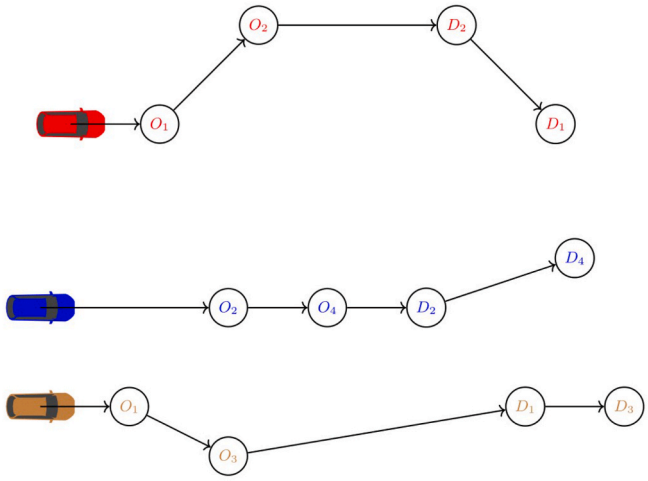

Fig. 2. Example of the Better-matching Effect. Both in the top row (low demand) and in the bottom row (high demand) we exhibit groups of size two. In the top row, the red vehicle is instructed to serve passengers 1 and 2, which are also marked with a red color. When the demand grows (bottom row), new passengers 3 and 4 appear, allowing the system to form more efficient groups. User 1 is now grouped with user 3 and served with a brown vehicle. Users 2 and 4 are grouped together to be served by a blue vehicle. The color of the passengers marks which vehicle serves them. Total delay decreases for the two users that remain from the top row, improving their perceived quality of service.

This effect also constitutes a source of scale economies for operators, as reducing detours also implies a reduction in VHT and VKT. The Better-matching effect is more evident when the assignment is done by batches of users (see Section 3.1), as in that case groups are formed all at once; however, it is also present under event-based approaches, as in that case the groups get formed sequentially as the individual users are assigned (for instance, in the example shown in Fig. 2, it is not relevant if the assignments were decided all together or one-by-one).

3.4. The Mohring effect

In public transport, the Mohring effect refers to the reduced waiting times that result from an increase in the optimal fleet size as a response to a greater demand (Mohring, 1972). Similar effects have been found for non-shared modes, such as taxis (Arnott, 1996) and ride-hailing (Castillo et al., 2017); recent papers by Kaddoura and Schlenther (2021) and Lee et al. (2021) have also found an analogous phenomenon in ODRP. Mathematically, this is represented by noting the obvious fact

that $B'(Y) > 0$ (more users require more vehicles), and that:

$$\frac{\partial t_h}{\partial B} \leq 0 \text{ for } h = w, a \tag{4}$$

Note that Eq. (4) includes walking (access) as well as waiting. This is because a larger fleet implies that vehicles are denser in space, hence shorter walks are required (when the system does not offer a door-to-door service), as also analysed in the case of public transport (Fielbaum et al., 2020b) and bike-sharing systems (Jara-Díaz et al., 2022). In the context of public transport, this is described as the "spatial counterpart of the Mohring Effect" by Fielbaum et al. (2020b).

3.5. Combination of the three effects

We have described three sources of scale effects: two positive ones (Better-matching and Mohring) and one negative (Extra-detour) in the case of users' costs. The Better-matching effect is induced directly by the greater demand, which enables the creation of more efficient groups, the Mohring effect is indirectly induced by a larger fleet, which diminishes waiting and walking times, whereas the Extra-detour effect is derived indirectly from the larger passenger occupancy rate of the vehicles, which increases detours and degrades the quality of service for some passengers. As the three effects might appear at the same time, it is uncertain which of them predominates. Formally, $\frac{\partial C_U}{\partial Y} = \frac{\partial C_U}{\partial \rho}\rho'(Y) + \frac{\partial C_U}{\partial B}B'(Y) + \frac{\partial C_U}{\partial Y}$, where the first term is negative and the other two terms are positive. For operators, if vehicle size is exogenous, both the Extra-detour effect and Better-matching effect reduce average costs, and therefore are sources of economies of scale; otherwise, the Extra-detour effect can lead to larger vehicles which increases operators costs as per Eq. (1). In Section 4, when running simulations, we identify the relationship between the scale three effects for different demand levels.

4. Numerical simulations

We now run numerical simulations of an ODRP system to analyse the occurrence of the scale effects discussed above, and to see which of them predominates depending on the circumstances. The system we simulate admits walks, is non-profit and adapts the fleet to have no rejections. To do so, we leverage the assignment method from Fielbaum et al. (2021) and Alonso-Mora et al. (2017), by deciding in real-time how many vehicles should be in operation, apart from the vehicle assignment and user assignment. In what follows, we discuss how to compute the endogenous fleet depending on the demand.

4.1. Computation of the number of vehicles in the ODRP system

In order to compute the fleet size together with the assignments between vehicles and users, we build upon the ODRP model proposed by Fielbaum et al. (2021). Such a model extends the one by Alonso-Mora et al. (2017) by optimising the pick-up and drop-off points, which might differ from the actual origins and destinations of the users when asking them to walk increases overall efficiency. Both models determine how to operate a fixed fleet of vehicles to serve the emerging requests. We extend these works by computing the fleet endogenously. We first explain briefly how the original methods work, and then describe this extension.

The ODRP system operates over a directed graph $G = (N, A)$. Each request $r = (o_r, d_r, t_r)$ is a triplet, representing the origin, the destination, and the time in which the trip is requested. Both the origins and the destinations are assumed to be placed over the nodes of the graph. The assignment model works using a *receding horizon* approach, meaning that it accumulates the requests that emerge during a fixed amount of time δ and assigns them all at once (hence it is batch-based), which updates each vehicle's route. When such an assignment is decided, the vehicles follow their updated routes, and the system begins to accumulate requests for a time δ again, starting a new iteration.

Let us focus now on a single iteration, denoting by R the set of requests to be assigned, and by V the current state of the fleet of vehicles. Each vehicle v is characterised by its position P_v and the set of requests assigned to it S_v (either in the vehicle or waiting for it). The assignment between R and V takes place following these three steps:

- Determine which are the feasible *trips*. A trip T is defined by a group of requests $req(T) \subseteq R$ and a vehicle $veh(T)$, so that T is feasible if the requests in $req(T)$ can be transported together by $veh(T)$, respecting some bounds on waiting and walking times, and on total *delay* (denoted, respectively, Ω_w, Ω_a, and Ω_v). Such bounds affect users in $req(T)$ and also in $S_{veh}(T)$, whose routes might be updated due to the new requests. The delay is defined as the extra time faced by a user compared to beginning her trip immediately, with no walking and following the shortest path between her origin and destination. Each trip T might be served by more than one route so that taking the route π imposes a cost to the system given by Eq. (5). The route π is defined by the nodes in which the vehicle stops to serve everybody, thus it contains implicitly the pick-up and drop-off points for every user.

$$cost(T, \pi) = \sum_{r \in reqt(T)} C_U(T, \pi) + \sum_{r \in S_{veh}(T)} \Delta C_U(T, \pi) + \Delta C_O(\pi) \tag{5}$$

Where the first term represents the users' costs for passengers in trip T, defined as a weighted sum between waiting, walking, and in-vehicle times; the second term represents the extra costs induced to the users that were being served by the vehicle prior to this assignment (because their waiting and in-vehicle times can increase); and the third term expresses the increase in operational costs, that are assumed to be proportional to the route length. The route that offers the minimum cost is selected, so that the trip T is characterised by a single figure that we denote $cost(T)$.

It is worth commenting that computing all the feasible trips can be computationally expensive, as their amount can increase exponentially with the number of requests (note that this increase in the number of feasible trips is the mathematical expression of the Better-matching effect, while the specific appearance of trips with many users represents the Extra-detour effect). Such an issue is faced first by making a smart search of the feasible trips (using that if vehicle v is able to serve group G, then it must be true that v can serve every subset of G as well), and also by using a number of heuristics, explained in detail by Fielbaum et al. (2021), to compute the sequence in which the users are served and the pick-up and drop-off points.

- Once the set Γ of potential trips is known with their respective costs, some of them are selected and constitute the actual assignment. To do this, an Integer Linear Programing (ILP) problem defined by Eqs. (6)–(8) is solved:

$$\min_{x,z \in \{0,1\}} \sum_{T \in \Gamma} x_T cost(T) + \sum_{r \in R} p_{ko} z_r \tag{6}$$

$$\text{s.t. } z_r + \sum_{T: r \in req(T)} x_r = 1 \forall r \in R \tag{7}$$

$$\sum_{T: veh(T) = v} x\_T \leq 1 \forall v \in V \tag{8}$$

Binary variables x_T represent the trips that are going to be executed (marked by $x_T = 1$). In the original model we are now describing, that operates with a fixed fleet, it is not always possible to serve all the trips (the number of vehicles might not be enough), so rejected requests are marked by $z_r = 1$. Each rejected request imposes a penalty p_{ko} to the system, so Eq. (6) is the objective function to be minimised when deciding the assignment. Eq. (7) ensures that each request is either rejected or belongs to a trip that is going to be executed, while Eq. (8)

ensures that each vehicle is assigned to no more than one trip.

- Finally, a rebalancing step instructs idle vehicles (i.e., those with no requests before the assignment and who did not receive anyone here) to move to certain areas where more vehicles are needed. In our setting, we execute a simple rebalancing step when simulating a feeder ODRP system, as explained in Section 4.2.1. We do not rebalance vehicles in the other scenarios, as the rebalancing step proposed by Alonso-Mora et al. (2017) sends vehicles to the places where users have been recently rejected, which does not occur here where we impose that everyone must be served.

In this paper, we extend this model to decide how many vehicles to use at the same time as deciding the vehicle and user assignments. To do so, we assume that the system begins with no vehicles, and that there are some spots in the city (which is a set of nodes $M \subset N$) where potential vehicles are placed. At each iteration (i.e., each time a batch of requests is assigned), the fleet of vehicles is composed of two sets: the one inherited from the previous iteration, plus a set containing one *non-activated* vehicle per request $r \in R$, that is located in the node in M that is closest to its origin o_r. If a non-activated vehicle is assigned to a group of requests, an activation cost c_A has to be paid, and the vehicle becomes available for the rest of the period of operation without paying the activation cost again. The parameter c_A includes all the costs that do not depend on the distance driven by the vehicle, such as capital costs. This is formalised by altering the cost of the trips. Denoting by $A(v) = 1$ if vehicle v is activated (i.e., inherited from a past iteration) and $A(v) = 0$ if not, Eq. (5) is modified to build the new cost function $cos\ t_A(v)$, given by

$$cos\ t_A(v) = cost(v) + c_A \cdot [1 - A(veh(T))] \tag{9}$$

Following Tirachini and Hensher (2011) and Jara-Díaz et al. (2017, 2020), we assume that both components of operators' costs grow linearly with the capacity of the vehicle. That is, recalling that K is the capacity of the vehicles, and denoting by c_O the proportionality constant that defines the costs depending on the routes' lengths, then:

$$c_O = c_{O1} + c_{O2}K, c_A = c_{A1} + c_{A2}K \tag{10}$$

As we now have one non-activated vehicle per request, it is always feasible to serve everybody. Therefore, we do not longer include variables z_r in the ILP to be solved, removing the second term from Eq. (6), and modifying Eq. (7) accordingly to ensure that each request belongs to exactly one assigned trip, i.e.

$$\sum_{T: r \in req(T)} x_r = 1 \forall r \in R \tag{11}$$

The problem is solved using the standard commercial ILP solver Gurobi. Finally, we include yet another extension to the base model: we assume that a fixed time τ is spent each time the vehicle stops to pick up or drop off one or more passengers. We include this fact because it is relevant when analysing scale economies, as sometimes the vehicle might use a single stop for more than one pick-up/drop-off, saving some time.

4.2. The scenarios

The analysis of scale economies requires increasing the demand level without changing its spatial distribution. We utilise the model described in Section 4.1 under two different types of scenarios. The first one is an ad-hoc network, namely an extension of the single-line model that has proved useful for scale analysis in transport systems in the past (see Section 2.2). The second one adapts a real-life database from Manhattan for this purpose.

4.2.1. Extending the single-line model

The traditional single-line model studies the operational characteristics of a public transport system in which the vehicles follow a predefined path, so everything is one-dimensional. Specific versions are:

- The circular model, in which the line tours a circuit that presents the same average number of users at every point. This model represents a line that carries a similar load all along its length.
- The linear model, in which vehicles travel in both directions along a linear corridor between two terminals. A particular case of the linear model is the feeder model, in which users board the vehicle across its path, and they all alight at the end. This model represents a line that goes to some relevant final destination, typically a public transport station, to board a high-capacity public transport mode (e.g., rail, Bus Rapid Transit).

In any of these alternatives, the vehicle route is fixed beforehand and always the same. We aim to extend this model, keeping most of its simplifying assumptions that make it a powerful tool, but allowing for online decisions regarding the routes. To do that, we deploy a grid surrounding each bus stop, where exact origins and destinations are situated. In the traditional model, such a grid can be seen as an underlying street pattern that does not need to be explicit because users need to walk towards the (fixed) bus stops anyhow. In such a case, walking times and distances are assumed exogenous, meaning that the operation and optimization of the public transport line are not affected.

To be precise, we assume that each bus stop belongs to a *zone*, which is an $a \times b$ grid, with a, b odd numbers, so that the bus stop is located at the centre of the grid. That set of stops represent where the potential vehicles for the ODRP system are located (the set M defined above). The central streets of the grid are bidirectional, and vehicles tour them at velocity v_1, whereas the rest of the streets are unidirectional,[5] with alternate directions and velocity v_2, where $v_2 < v_1$. Having streets of different velocities and directions help to capture that not all routes are equally good for the vehicle to follow. The whole network is formed by chaining consecutive zones. If there are Z zones, this makes a $Z{\cdot}a \times b$ grid in the feeder model; in the circular model, the same happens, but the last zone is chained with the first one, forming a circular grid. Both networks are depicted in Fig. 3.

Regarding the demand, we want to keep the homogeneity assumptions from the single-line model but enabling for more complex routes. A constant number of users Y emerge per time unit, and the exact origin is random: we first choose the zone with uniform probability; within that zone, the central node is chosen with probability p, the rest of the nodes located in the central streets with probability $p\gamma$, and the nodes out of the central streets with probability $p\gamma^2$. The parameter p is adjusted to make the sum of the probabilities within every zone equal to 1, and the parameter $\gamma \in (0, 1)$ controls how dispersed the demand is within a zone (the lower the γ, the more concentrated the demand in the vicinity of the bus stop). The destination is computed differently depending on the model: in the feeder one, everybody goes to the centre of the final zone, whereas in the circular model, the destination zone is located l zones ahead, plus a random variable that is obtained rounding a normal distribution with mean zero and variance σ^2; the exact destination is found within that zone using the same rules involving p and γ as for the origin.

As mentioned above, in the feeder model we need to rebalance idle vehicles to prevent them from accumulating in the common destination of all users: after reaching that node, they are sent towards the central node of the first zone (i.e., the one located at the largest distance from the shared destination). Such vehicles will not necessarily arrive there because they will be considered available in the following iterations, meaning that they might receive new passengers before reaching the

[5] In the feeder model, the first and last transversal streets are also bidirectional so that there are no isolated nodes.

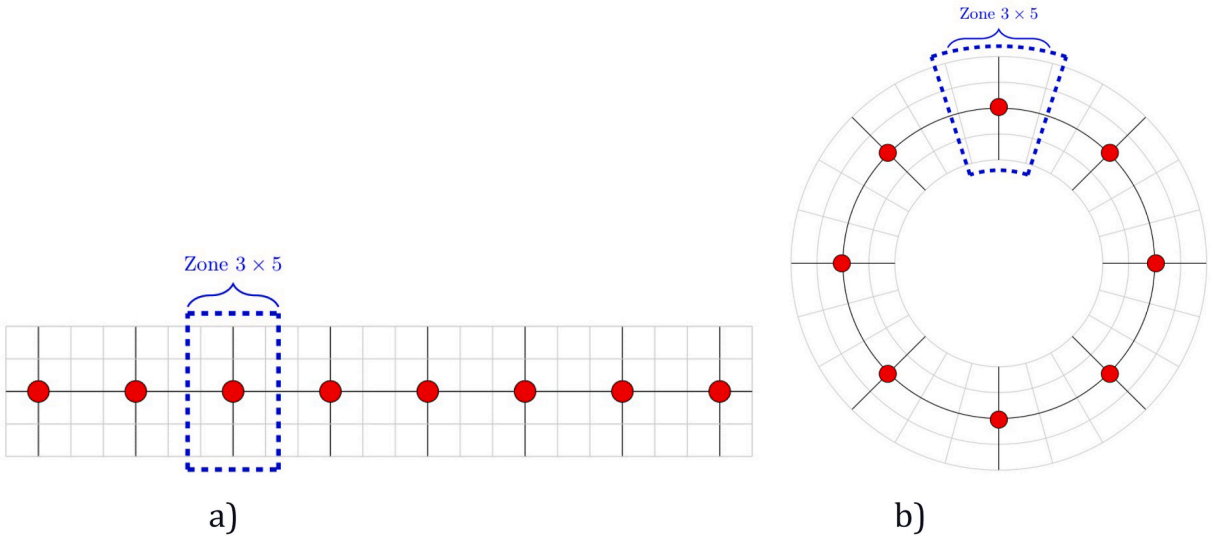

Fig. 3. Extensions of the single-line model to recreate the network in which the ODRP system operates, replacing either a feeder line (a) or a circular line (b). Origins can be placed in any intersection, and the same happens with destinations in the circular model. In both cases, there are 8 zones, each formed by a 3x5 grid. Red dots represent the stations in which the ODRP vehicles begin their journeys. Dark black streets are bidirectional and can be toured with a higher speed. The traditional single-line model is recovered by considering only the long avenue that connects all the red dots.

first zone.

4.2.2. The Manhattan scenario

Similar to other studies that have simulated ODRP (e.g., Alonso-Mora et al., 2017; Simonetto et al., 2019), we leverage the public database generated by the NYC Taxi & Limousine Commission, that contains all the taxi trips that are executed in Manhattan. For each trip request, we know the number of users, its time, origin, and destination. The graph is composed of 4092 nodes and 9453 arcs. We consider 1 h of the operation of the system, just after 1 p.m. of January 15th, 2013, which has 10,774 trip requests. In order to perform scale analysis while keeping the spatial distribution of the demand somewhat constant, we repeat the simulations several times, each time increasing the number of requests that are considered. To be precise, for every repetition, and before running the simulations that represent 1 h of operation, we decide randomly whether each request is considered or not. The probability of including a request is the same for every request, and this probability increases for each repetition. By this means, the set of requests to be assigned is larger every time, but their spatial distribution is kept on average, as the requests are selected randomly from the same universe set.

Here there is no natural candidate for where the vehicles should begin their journeys when they are activated (the equivalent to the centre of each zone in the single-line model). To face this issue, we cluster the network, finding the minimum number of centres such that every node in the network can be reached in less than Ω_w from at least one centre. This problem is solved by an ILP (described in detail by Wallar et al., 2018), which leads to 19 centres. Therefore, when a request needs to travel from a node x, the corresponding potential vehicle will be placed in the centre that is located closest to x.

4.2.3. Definition of the bounds in the quality of service

As explained in Section 4.1.1, the assignment procedure in ODRP imposes predefined bounds on the quality of service, namely maximum waiting (Ω_w) and walking (Ω_a) times, as well as a maximum total delay[6] (Ω_v). Defining such bounds is a relevant issue, as it has relevant impacts on the performance of the ODRP system. For instance, if the bounds are too tight and users are too spread, then the system might require to allocate almost one different vehicle per request, leading to a huge fleet; on the other hand, if the bounds are too large (or inexistent), one single vehicle might be able to serve all the requests, but offering an awful (and unrealistic) quality of service. We will consider two ways in which these bounds are defined:

4.2.3.1. Endogenous bounds. First, we consider a case in which the bounds are calculated as a function of the demand, using longer time windows when the demand is low. This is done for the single-line model, as it mimics what passengers usually face when using public transport: when they want to make a trip on a high-demand corridor, they can rapidly find a bus (or any alternative mode they are using), and the contrary happens in low-demand areas (a similar argument has been proposed by Yan et al., 2020 when proposing their *dynamic waiting strategy*). Thus, we define the bounds to replicate this behaviour, by means of the classical single-line model by Jansson (1980) and the posterior adaptations by Jara-Díaz and Gschwender (2009), described in Appendix A.1, where the key variable is the optimal frequency f. The bounds are defined as follows:

- **Waiting:** The maximum waiting time that can be faced in the public transport system occurs when a passenger arrives at the station just after a bus leaves, waiting for $1/f$ (a quantity that decreases with the number of passengers). Recalling that when a vehicle is activated, it goes from the station to the pick-up point, we need to ensure that there is always enough time to wait for such a movement. Denoting by t_1 the vehicle-time from the station to the corner of the zone's grid, we use $\Omega_w = max\left\{\frac{1}{f}, t_1\right\}$.
- **Walking:** The maximum amount of walking in the public transport systems is t_2, defined as the walking time between the station and a corner of the zone's grid, so we use $\Omega_w = t_2$. When we simulate the case in which ODRP offers a door-to-door service, this bound is reduced to zero.
- **Delay:** There are two sources of delay in public transport with respect to the time in the vehicle: walking and waiting. The first one should be accounted for twice, at the origin and destination. Therefore, we use $\Omega_v = max\left\{\frac{1}{f}, t_1\right\} + 2t_2$.

[6] Such bounds ensure that users will indeed accept the assignment proposed by the system rather than searching for an alternative mode. Moreover, without them the algorithmic burden of the problem would be unmanageable, as every possible group of users could be feasibly served by any vehicle.

4.2.3.2. Exogenous bounds. Defining the bounds as a function of the demand volume may have a drawback, namely that the varying bounds may interact with the other scale effects that we are analysing. For this reason, we also study the case in which the bounds are exogenous, using $\Omega_w = 3$ minutes, $\Omega_a = 4$ minutes, and $\Omega_v = 6$ minutes.

4.3. Results

We simulate 1 h of operation of the ODRP system, for increasing demand levels, in order to identify scale effects. The numeric values of the parameters are shown in Table A.1 in the Appendix. All figures in this section use a logarithmic scale in the x-axis, because the phenomena that we study tend to stabilise when the number of passengers is high, so zooming in the lower values helps the analysis. The simulations are run for different values of K (the size of the ODRP's vehicles), including vehicles with capacity for 2, 3, 4, and 5 passengers.

Most results consider the base case in which we assume the availability of automated vehicles (AV) and walks are allowed. We assume that AV differ from human-driven vehicles in the parameters that represent operator costs: on the one hand there is a reduction in operating cost due to savings in driver wages, on the other hand there is an increase in capital cost due to the added cost to provide vehicles with automation capabilities, which is taken from Tirachini and Antoniou (2020). We will assume that the velocity at which vehicles run do not depend on such a technology: differences in velocity due to automation are still uncertain, as AV might run faster (thanks to better coordination among vehicles) or slower (due to safety reasons, particularly in urban roads when surrounded by pedestrians, cyclists and human-driven vehicles). Specific assumptions about differences in velocity may have a large impact on the results (Tirachini and Antoniou, 2020).

4.3.1. Time windows adapted endogenously to demand - circular model

We first describe the results when the time windows $\Omega_w, \Omega_a, \Omega_v$ are calculated as a function of the demand. We show the results of the circular model in Figs. 4–8. Fig. 4a shows a condensed way to describe the quality of service of the ODRP system from the users' point of view: total delay, i.e., the extra time faced by them when they use this system instead of travelling in a private vehicle. Fig. 4b exhibits the average occupancy rate [pax/veh] per vehicle ρ at the end of the simulation, which was identified as the crucial factor for the existence of diseconomies of scale (the Extra-detour Effect). Total delay includes walking time, waiting time, and detour once in the vehicle. Scale effects are evident: At the very beginning of the curve, up to around 250 passengers/h, there is a reduction in total delay. However, when the number of passengers continues to grow, diseconomies of scale appear as the average delay increases to 5 min/passenger for demands up to almost 1000 passengers/h. Remarkably, the appearance of diseconomies of scale coincides exactly with the moment in which ρ starts to increase (before that, it slowly decreases, which is explained by the changes in the time windows). Then, the average delay is once again reduced, to reach around 2 min/passenger for 3000 passengers/h.

To delve into the curves from Fig. 4 and identify the emergence of the three effects described in Section 3, we disentangle the total delay per passenger in its three components in Fig. 5: Waiting (a), walking (b), and detour (c). Waiting times evolve similarly to total delay.

Let us begin our analysis after the strong drop at the beginning of the graph. The remainings of the curves reflect that the average delay first increases and then slowly decreases. Diseconomies of scale emerge when Y reaches about 250 passengers/h. Until that point, there is little sharing in the system (below 1.3 users per vehicle), because it is difficult to find compatible users, implying that most users travel alone.[7] When vehicles begin to be shared with more people, one of its consequences is that vehicles do not go directly to pick up the users but deviate to serve some co-travellers, hence increasing waiting times. This effect dominates for demands greater than 250 passengers/h. The same phenomenon can be seen related to walking and the detour, which also start to increase when crossing the same threshold. Noteworthy is that the smaller the vehicle, the lower the detour, and that detours can be negative, meaning that the distance between the pick-up and drop-off points might be lower than between the corresponding origins and destinations (due to walking). Therefore, our simulations confirm the Extra-detour effect as **a relevant source of diseconomies of scale in ODRP systems: an increase in the number of users implies that the vehicles will be shared by more passengers, which increases average travelling times.**

The Extra-detour Effect eventually gets exhausted. At some point, the vehicles no longer increase their load (when they are running at capacity, considering their current passengers and the ones that are waiting to be picked up). When this happens, Fig. 5a and b reveal that waiting and walking times begin to diminish. That is to say, the two sources of scale economies described in Section 3 begin to dominate: the Mohring Effect and the Better-matching Effect. We can synthesise these scale effects by stating that **two relevant sources of scale economies in ODRP are that the increase in the number of users leads to 1) a larger fleet, which reduces waiting and walking times, similar to the Mohring Effect in fixed-route public transport, and 2) matching users whose routes are more compatible.**

The quick drop at the beginning of the curve is explained by the Mohring Effect, but only regarding waiting times because for such a demand there is almost no walking. As vehicles' occupancy rate does not increase yet, neither the Extra-detour nor the Better-matching Effects operate significantly. Actually, the average occupancy rate decreases slightly, due to the decrease in Ω_w and Ω_v, which makes the Mohring Effect even stronger as more vehicles are needed.

It is worth commenting that the Mohring Effect is usually more important at low demands because when the number of vehicles is already large, the marginal impact of an additional vehicle is low in reducing waiting times. Then, the drop in average times at the end of the curve is mainly driven by the Better-matching Effect.

The fact that there is no walking at all when the demand is very low is explained because there is little sharing and it is difficult for vehicles to chain consecutive trips. Thus, many times the only user involved when deciding a vehicle's route is the one being transported, and for her it is more comfortable to have a door-to-door service (as usual in the literature and shown in Table A1 in the Appendix, we assume that $p_a > p_v$). The absence of walks also explains why the detour is longer at the beginning of the curves. This suggests that if p_a was low (but greater than zero), the Mohring Effect would affect walking times for low demand volumes as well. Note that the average walking time increases slightly before the large jump at 250 passengers/h, which is explained by the changes in the time windows, as sometimes such short walks can enable a group that would be otherwise infeasible.

The analysis above confirms the three sources of scale discussed in this paper. However, the varying bounds on the quality of service do play a role in the analysis, which is why in Section 4.2.3 we analyse the

[7] This occurs in some real-life scenarios. In several towns around Munich, Germany, during some specific time windows in which the demand is very low (between 17:30 and 5:45 in working days), the public transport agency sends private taxis to fulfil it. This result has also been reported by Daganzo et al. (2020) when comparing three different demand levels with vehicles of capacity 2 pax/veh.See https://www.mvv-muenchen.de/mobilitaetsangebote/bedarfsverkehr/mvv-ruftaxi/index.html (Accessed: 10/08/2022).

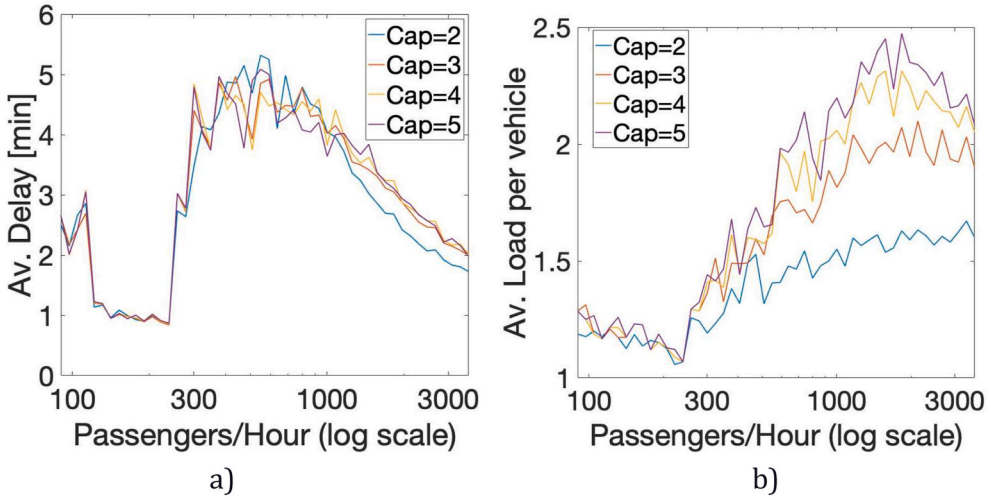

Fig. 4. a) Average total delay faced by the users, and b) Average load per vehicle at the end of the simulation, in the ODRP system for the circular model with endogenous time windows, as the number of hourly passengers grows. Different curves represent different vehicles' sizes.

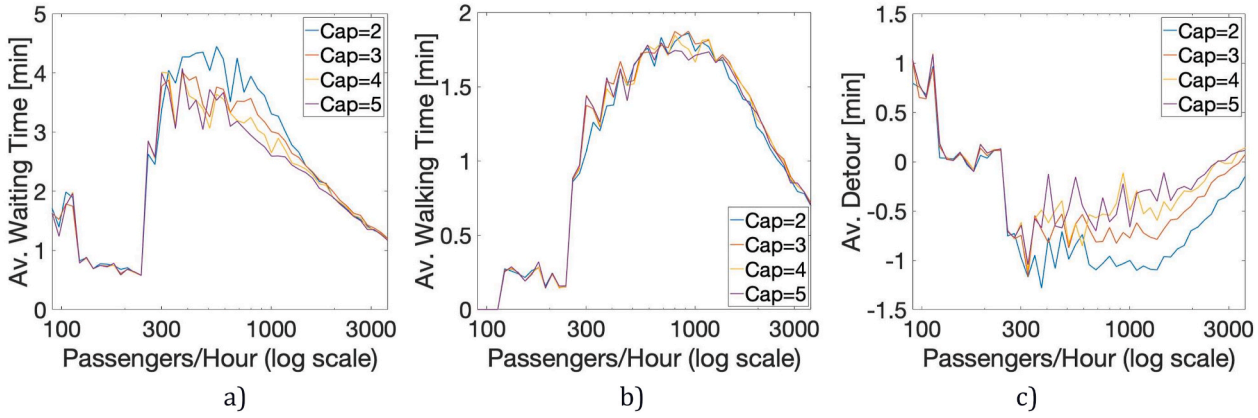

Fig. 5. Average waiting time (a), walking time (b) and detour (c), faced by the users of the ODRP system in the circular model with endogenous time windows, as the number of hourly passengers grows. Different curves represent different vehicles' sizes.

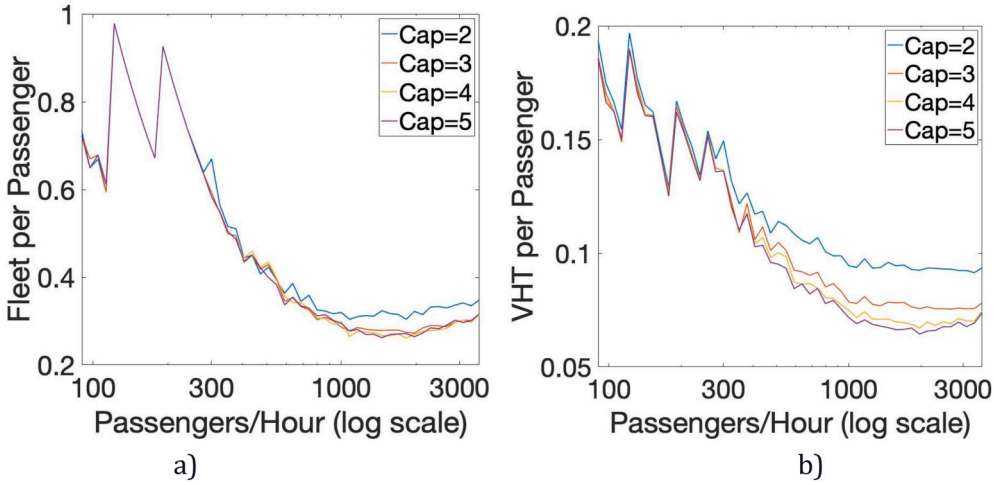

Fig. 6. Fleet size (a) and Vehicle-Hours-Travelled (b), normalised by the number of passengers, as this last quantity grows. Different curves represent different vehicles' sizes.

case with fixed bounds, where we show that the qualitative conclusions remain valid.

The comparison among different vehicle sizes is also informative. The smaller the vehicle, the lower the number of passengers per vehicle, and thus the detour. Walking times are not affected significantly by the vehicle capacity adopted. On the other hand, waiting times are slightly

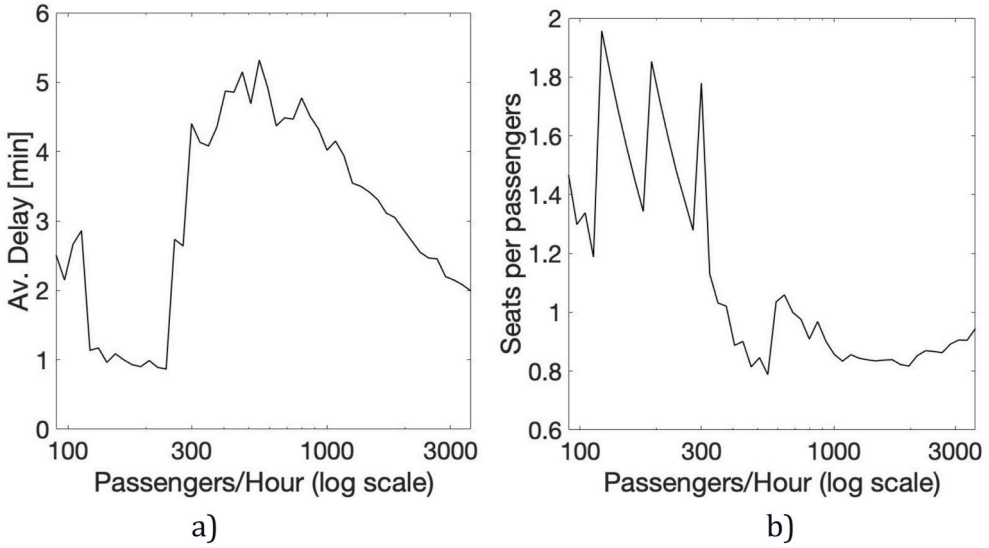

Fig. 7. Average delay (a) and Seats per passenger (b), yielded by the ODRP system in the circular model with endogenous time windows, as the number of hourly passengers grows, when the optimal capacity is selected.

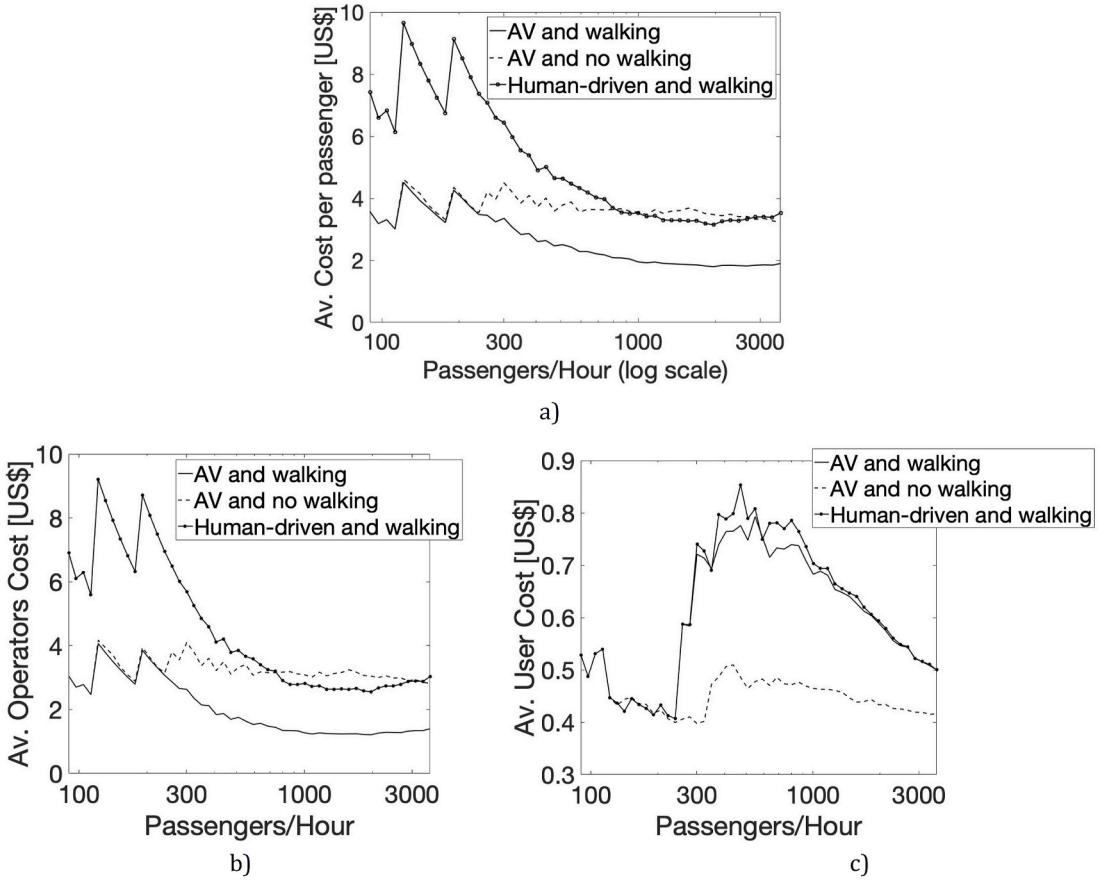

Fig. 8. Average (a) total costs, (b) operator costs and (c) users cost per passenger in the circular model with endogenous time windows, as the number of hourly passengers grows, when the optimal capacity is selected. Different curves represent different types of vehicles and whether walks are enabled in ODRP.

larger for smaller vehicles when the demand lies in the range of 250–1000 passengers/h. As the fleet size is mostly unaffected by the vehicle capacity within that range (see Fig. 6a), it is more likely that the assigned vehicle is not immediately available when vehicles are small.

The evolution of the components of operator costs (depicted in Fig. 6) is mostly characterised by scale economies when exceeding the threshold in which vehicles start to be shared more intensively; before the threshold, it exhibits an irregular pattern in which the randomness of the requests play the most relevant role. This is reflected in the fleet size (Fig. 6a), which also exhibits scale economies in public transport, and too in operating costs (vehicle hours travelled VHT, Fig. 6b). That is to say, **an operator-related source of scale economies is given by the**

increase in sharing, which makes the number of vehicles and VHT increase less than linearly with the number of users. It is noteworthy that using smaller vehicles requires a larger fleet when used at capacity, which also increases VHT. Both curves eventually stabilise, meaning that this source of scale economies gets exhausted.

So far, we have exhibited results for a range of vehicle capacities, from 2 to 5 passengers/veh. However, the system should utilise vehicle sizes that minimise total costs. Our results indicate that the smallest vehicles (capacity 2) should be used if $Y \leq 550$, and capacity 3 thereafter. Fig. 7 synthesises scale effects for users and operators when the capacity is optimised. The delay curve (Fig. 7a) looks almost exactly as Fig. 4, meaning that all the scale phenomena discussed above remain valid. Fig. 6 implies that both the number of vehicles and VHT still exhibit scale economies when the capacity is optimised, but there remains one aspect to be analysed: the number of seats S, defined as the product of the number of vehicles and their capacities. Recall that, according to Eq. (10), operators' capital and operating costs depend both on the total number of vehicles and on S. The evolution of S when the capacity is optimised is shown in Fig. 7b: it is similar to what we observed regarding fleet size (first erratic and then scale economies), but with a small jump when the optimal capacity switches from 2 to 3 (around 600 passengers/h in Fig. 7b). We note that in real-life implementations, the fleet might not be homogeneous, i.e., it is possible to have vehicles of different sizes, which would decrease the size of these jumps as the average K would approach a continuous function; however, optimally routing and operating a heterogenous fleet is quite complex, let alone designing it, so this is beyond the scope of this paper.

In Fig. 8, we synthesise the results by depicting average total costs, and also average users and operators costs. We further include two alternative scenarios: forbidding walks (i.e., providing door-to-door service), and utilising human-driven vehicles instead of AVs, which diminishes capital costs but includes the drivers' wages. Fig. 8a shows the average cost per user: in all three scenarios, we observe the same situation, namely, no clear trends for very low demands and economies of scale after a certain demand threshold is reached. This implies that the sources of diseconomies of scale that we identified for the users get outweighed by the sources of economies of scale for the operators, **leading to a global situation of economies of scale that eventually get exhausted**. Note that having overall economies of scale means that the sources of diseconomies of scale (crucially the Extra-detour Effect) can be compensated through pricing. Some relevant insights can be obtained from the comparison between the different scenarios and vehicle technologies:

- Using AVs reduces the total cost to a considerable extent. This fits intuition, as having drivers for each small vehicle can increase total costs significantly (Bösch et al., 2018).
- In general, operator costs (Fig. 8b) are larger than user costs (Fig. 8c) and the shape of the total cost curves is mostly driven by the shape of the operator cost curves. In Fig. 8 c, sharing a vehicle significantly increases users' cost due to extra waiting, walking (when admitted), and detours. Users would prefer not to walk, but because walks are short, the advantage of a door-to-door service is much smaller than what is gained in terms of operators' costs when walks are admitted.
- When the number of users is large, enabling walks can be as important as changing the vehicle technology: both non-solid curves exhibit similar values of average total cost in Fig. 8. **In fact, an ODRP system with human-driven vehicles that enables walking has a lower total cost than a system with AVs without walks, for some demand levels.** This is a remarkable finding regarding the value of designing an ODRP system with short walks.
- On the other hand, as there is little walking when the number of users is low (the system works similar to a private door-to-door service), for demands below 250 passengers/h the corresponding impact of enabling walks is negligible.
- The curve in which walking is not allowed exhibits returns that are almost constant to scale (similar to the findings by Militão and Tirachini, 2021b). Therefore, **admitting walks happens to be crucial to trigger scale effects.**

4.3.2. *Time windows adapted endogenously to demand - feeder model*

There is an emerging research trend that studies the potential of ODRP services to help solve the so-called "last-mile problem", i.e., as a feeder that connects the main public transport stations with the specific origins (or destinations) of the users (e.g., Bürstlein et al., 2021; Chen et al., 2020; Fielbaum, 2020; Kim and Schonfeld, 2014; Leffler et al., 2021; Ma et al., 2019; Wen et al., 2018). For the ODRP system, the main difference with respect to the circular model is that everybody shares one extreme of the trip, which means that this model can also represent the case in which there is a very attractive destination, such as the city centre. In our simulations, all users are travelling to the same destination (for instance, to take a second vehicle that does not affect the ODRP operation). Therefore, compatible routes are much easier to find. The only requirement is that when a vehicle is following a route, new passengers have to be located close to that route. This demand pattern has a significant effect in the simulations: for the same number of users, the number of feasible trips is multiplied by about twenty compared to the circular model. This increases the computational burden significantly, which is why here we simulate only up to capacities equal to four.

There is yet another relevant difference related to idle capacity. As users move all in the same direction, and the network is no longer circular, the vehicles must actively return in order to find some new passengers. Recall that this is executed through a rebalancing step: idle vehicles are sent towards the other extreme of the network, but they might not arrive there because they are still considered available for the emerging users.

The results of the simulation are depicted in Fig. 9, considering the base model (AVs and enabling walks). Fig. 9a condenses the information regarding users' costs by displaying the average delay, which shows the same trends as observed in the circular model, verifying the presence of the three sources of scale economies discussed above. Fig. 9b shows the average number of passengers per vehicle (excluding vehicles being rebalanced), confirming that vehicles start to increase their occupancy rate when some threshold in the number of passengers is exceeded. Moreover, the usage of the vehicles is much higher than in the circular model. When looking into total costs (Fig. 9c), the same conclusions obtained for the circular model remain valid: average costs do not show a clear trend at the beginning, and scale economies prevail afterwards until they eventually get exhausted.

4.3.3. *Fixed time windows - circular model*

We now show the results when the bounds on the service times Ω_w, Ω_a, Ω_v are exogenous and independent of the demand level. This case requires very long computational times, as the high-demand scenarios do not have short time windows, which leads to combinatorial problems of large size. This is why we only present results from the circular model and consider vehicles with capacity 2 and 3.

Results are shown in Fig. 10, and reinforce the conclusions previously discussed. Fig. 10b is very illustrative, as it shows that for very low demand (lower than 30 passengers/h) there is no sharing at all (at the end of the simulations), so that the only relevant scale source is the Mohring Effect and average delay (Fig. 10a) decreases. When the average occupancy rate begins to increase (Fig. 10b), the same happens with the average delay, triggered by the Extra-detour Effect. For a demand larger than 100 passengers/hour, the average delay starts to decrease again, coinciding with the threshold after which the average load increases at a lower rate. Fig. 10c shows that when all costs are accounted for, scale economies prevail and eventually get exhausted.

4.3.4. *Fixed time windows - Manhattan*

We now show the outcome of simulating 1 h of ODRP operation over

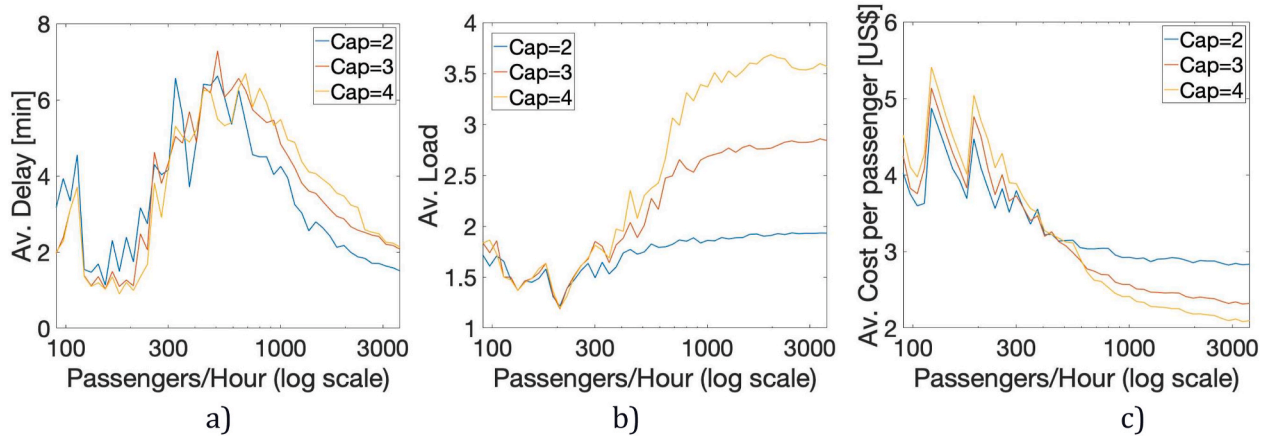

Fig. 9. Average delay (a), active vehicle's load at the end of the operation (b) and costs (c), faced by the users of the ODRP system in the feeder model with endogenous time windows, as the number of hourly passengers grows. Different curves represent different vehicles' sizes.

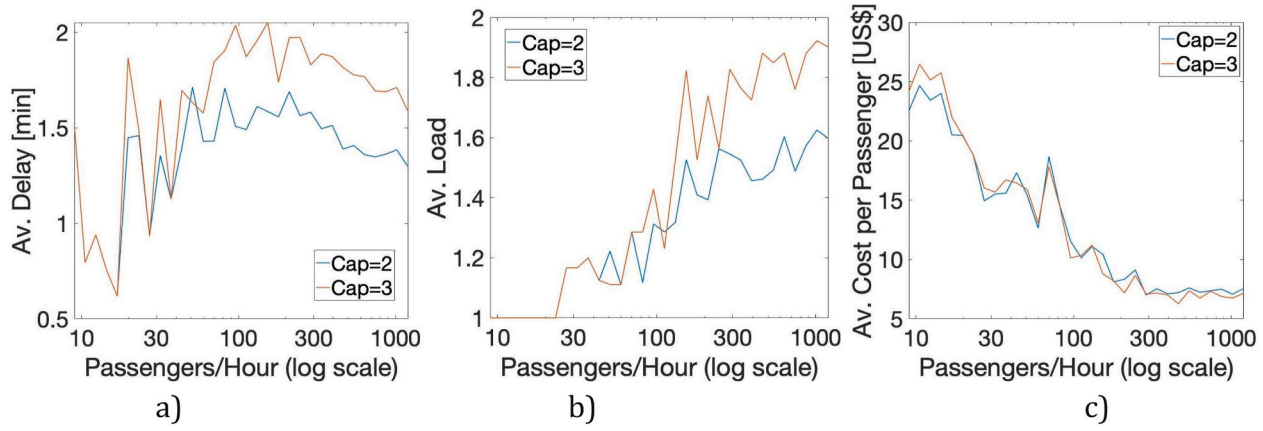

Fig. 10. a) Average total delay faced by the users, b) Average load per vehicle at the end of the simulation, and c) Average costs, in the ODRP system for the circular model with exogenous time windows, as the number of hourly passengers grows. Different curves represent different vehicles' sizes.

Manhattan, considering subsets of increasing size extracted from a real-life database of taxi travellers. Results are depicted in Fig. 11, and show that the qualitative sources of scale that were discussed in Section 3, and verified numerically in ad-hoc networks in the previous subsections, remain valid under this real-life scenario. The delay (Fig. 11a) first decreases (Mohring effect), then increases (Extra-detour effect), and then decreases again (Better-matching effect). The threshold where the trends change coincides with the changes in the average occupancy rate per vehicle (Fig. 11b): when vehicles begin to be shared, the delay starts to increase, and when the sharing rate becomes more stable, the delay decreases. Finally, Fig. 11c shows once again that when all costs are taken into account, scale economies prevail but eventually get exhausted.

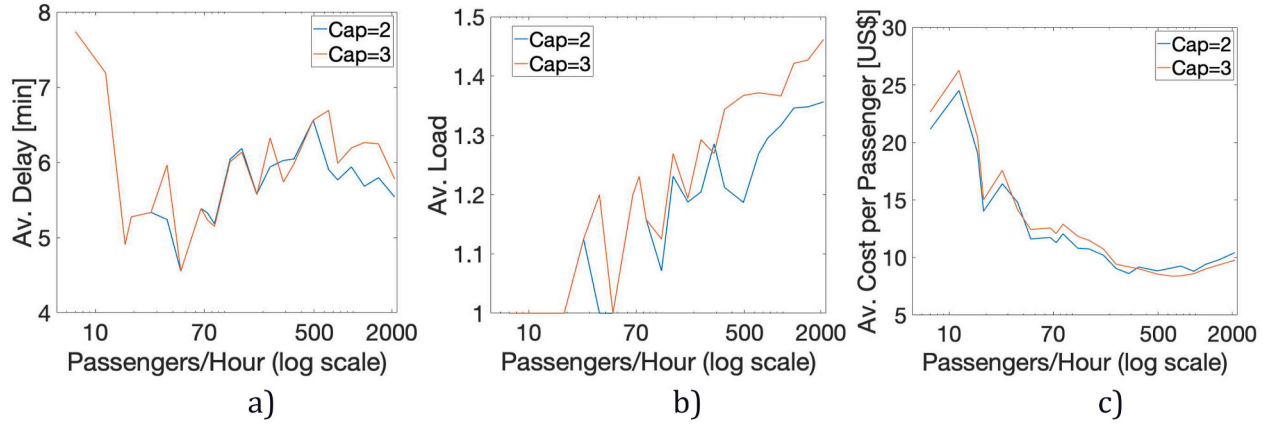

Fig. 11. a) Average total delay faced by the users, b) Average load per vehicle at the end of the simulation, and c) Average costs, in the ODRP system operated in Manhattan. Different curves represent different vehicles' sizes.

4.4. Evolution and predominance of the three user-related scale effects

We can now discuss the question posed at the end of Section 3: which of the user-related scale effects predominate and under which circumstances? The evolution of user-related scale phenomena as the number of users increases is described in Fig. 12. It is a stylized schematic figure that divides the analysis into three demand ranges, representing the respective demand segments (as seen in Figs. 6a, 9a and 10a, and 11a) in which the average delay first decreases, then increases, and then steadily decreases again. Fig. 12 shows the so-called *Degree of scale economies (DSE)*, which is formally defined for any production function as the ratio of the average cost to the marginal cost: this means that there is a threshold in $DSE = 1$ determining whether economies or diseconomies of scale prevail. The mentioned three sectors are:

- When the number of passengers is low (first segment of the curve, e.g. in the late-night period), users hardly share a vehicle, so that the Extra-detour and the Better-matching Effects are almost non-existent. This means that the Mohring Effect (which is more prominent when the demand is low) prevails, and there are economies of scale.
- Eventually, users begin to share the vehicle, and the system enters into the second demand range. The Extra-detour Effect begins to operate, and diseconomies of scale prevail. The Mohring Effect is still present, but dominated. The Better-matching Effect also starts to operate but mildly due to the increased passenger occupancy rate. The minimum of the curve represents the point at which vehicles' load increases at the fastest pace.
- Finally, when the vehicles cannot carry more passengers (they are full), the Extra-detour Effect disappears, and the Mohring Effect has little impact. The Better-matching Effect, on the other hand, is fully operative, leading to $DSE > 1$. Eventually, *DSE* converges to 1 as all these sources get exhausted.

4.5. Comparison with an idealised public transport model

As the single-line model resembles the operation of a traditional public transport line, it is natural to analyse under which conditions ODRP could replace such a line. A precise model of the public transport is out of the scope of this paper; however, we do perform a comparison with an idealised public transport line, whose frequency and bus capacity are optimised following a procedure described in Appendix A.1.

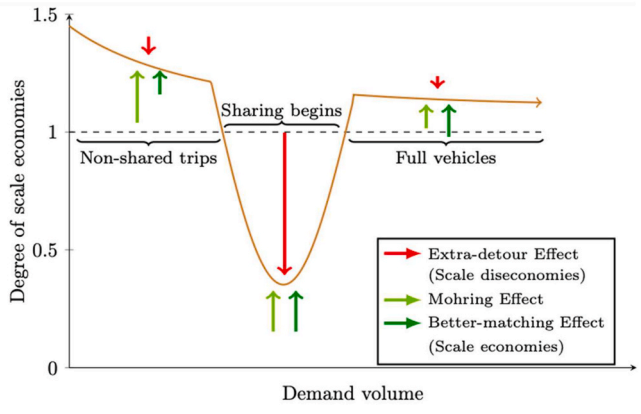

Fig. 12. Synthesis of the three sources of users-related scale effects for ODRP systems. The y-axis represents the degree of scale economies (*DSE*), so that scale economies prevail when $DSE > 1$ and the contrary happens when $DSE < 1$ (constant returns to scale if $DSE = 1$); the x-axis represents the number of users, and we do not include concrete numbers because this is a schematic representation. The direction of each arrow represents if it pushes *DSE* upwards (i.e., scale economies) or downwards (i.e., scale diseconomies), while its length represents its magnitude.

Such a comparison is depicted in Fig. 13, and is informative regarding the trends in the respective curves. Fig. 13 depicts the ratio between the total costs (including operators and users) of ODRP and public transport, considering both the circular and the feeder models. In ODRP, we select the capacity of the vehicles that minimises total costs, considering endogenous time windows. ODRP is in the numerator, so that a value lower than 1 implies that ODPR provides the lowest total cost. The most relevant conclusions of this comparison are the following[8]:

- ODRP should only be preferred if the demand is very low, in line with the findings of previous research efforts, as described in Section 2. This result is driven by the small size of the ODRP vehicles, and relates to the almost door-to-door scheme that results in such scenarios. This last characteristic also explains why ODRP is more competitive in the circular model for low levels of demand, as in the feeder model, public transport also has zero walking at the destination, softening the benefits of ODRP.
- For large demand levels, ODRP is more competitive in the feeder model. Note that in public transport, vehicles also need to "rebalance", i.e., to return empty to the other extreme of the network. In this case, all vehicles have to arrive there, as their route is fixed. In ODRP, they do not need to arrive at that extreme, so that flexibility plays a role in diminishing the idle capacity of the system.
- For large demand levels, curves tend to stabilise, which is a natural result of the constant returns to scale that characterises all these systems in such scenarios.

In all, if one has to choose between using only ODRP (with small vehicles) or only traditional public transport, the former should be chosen only for low-demand zones. However, our results regarding the presence of scale economies when the demand is large, suggest that other types of integration could yield even better results, utilising both systems in some complementary way to take advantage of the good quality of service that can be offered to the users. How to design such an integrated system is a broad question that goes beyond the scope of this paper, but recognizing that there might be room for improving public transport provision in high-demand zones by means of smart utilisation

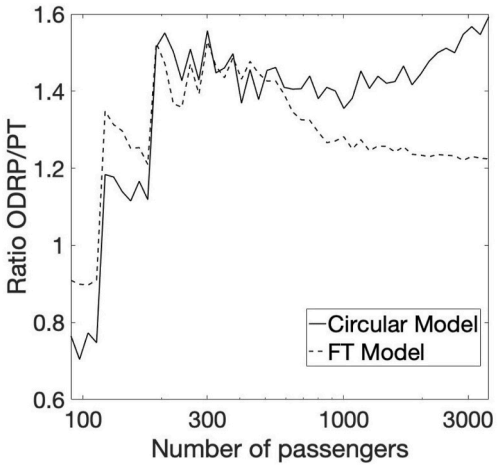

Fig. 13. Comparison between ODRP and public transport average costs as the number of hourly passengers grows, when the optimal capacity is selected, using AVs and enabling walks. Different curves represent the circular and the feeder model.

[8] Both problems might be faced with ad-hoc techniques like having some vehicles serving only the last portion of the line, i.e., a "short-turning" strategy, potentially combined with deadheading, as studied by Cortés et al. (2011).

of ODRP systems is a promising venue for further inquiry.

5. Conclusions and future research

In this paper, we have identified and analysed the sources of economies and diseconomies of scale in on-demand ridepooling (ODRP) systems. To do this, we have extended a state-of-the-art assignment method for ODRP, in order to optimise the fleet size together with the decisions of how to group the users and which vehicles are assigned to each group of passengers.

We have discussed three scale effects affecting users, implying both positive and negative externalities to the other passengers. Positive externalities are the Mohring effect, i.e., a reduction of waiting times as demand grows, and the "Better-matching effect", i.e., the reduction of access times, waiting times, travel times and operator costs that it is possible because more efficient groups of passengers can be formed when demand grows. The negative externalities relate to increasing the number of users per vehicle, which induces longer detours, a phenomenon we call the "Extra-detour Effect". There are only positive externalities on the operators' side, namely that vehicles can be used more intensely so that the fleet size grows less than linearly as a function of demand.

Such effects have been theoretically discussed and verified in simulations, which have been run considering several different scenarios, including two simplified single-line configurations (in which ODRP is assumed to operate in the equivalent of a zone covered by a public transport line) and one real-world network from Manhattan, New York. Results are remarkably similar across all the scenarios analysed. The simulations have enabled determining which of these scale effects prevail as the number of passengers increases. First, for low demand levels, there is little sharing and scale economies prevail thanks to the Mohring effect. Then, as total demand grows, users start to share rides and the Extra-detour effect dominates, leading to a global situation of diseconomies of scale for users. Finally, if demand increases even further, vehicles run at capacity and again positive effects prevail thanks to the Better-matching effect.

We have found that for the efficient operation of ODRP in a setting without request rejections, the possibility of asking the passengers to perform short walks to pick-up points is crucial to keep total costs low, both for users and operators. In particular, we have found that an ODRP system with human-driven vehicles and walks allowed has a total cost at a similar level to that of a door-to-door ODRP system with automated (fully driverless case) vehicles. This finding has significant implications for the current and future design of mobility systems based on shared vehicles and shared rides, either with human-driven or automated vehicles.

If the system designer has to choose between a traditional public transport line or an ODRP system, the latter should be mostly preferred for low-demand zones. However, the scale effects in ODRP suggest that there could be other ways of integrating both systems to enhance public transport and attract users from private modes in high-demand scenarios, especially for feeder-like systems. Understanding how this could be done is the most relevant future research question that emerges from this paper.

Our findings might be limited by the assignment method we utilise. However, even if another numerical setup may change the average costs estimated, we have qualitatively argued that (i) the three scale effects under scrutiny do exist in ODRP systems in general (including references to other studies when appropriate), (ii) they interact with each other and (iii) the influence of each other in pushing average costs up or down depends on the total demand level.

As extensions to the current approach, including some market-related effects is a promising path. Considering that the fleet is owned by one or more for-profit shared-mobility companies might have an influence on scale analysis that is worth studying. Similarly, it is worthwhile to consider the case where the supply is not centrally controlled, i.e. drivers can choose when to connect and which passengers to accept. Finally, users' strategic responses to different pricing policies can have a relevant effect on the degree of sharing and therefore on the Extra-detour and Better-matching effects.

CRediT authorship contribution statement

Andrés Fielbaum: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Alejandro Tirachini:** Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Javier Alonso-Mora:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Supervision, Resources, Project administration.

Acknowledgments

Part of this work was developed when the second author was affiliated to Instituto Sistemas Complejos de Ingeniería (ISCI) in Chile, financial support from ANID PIA/PUENTE AFB220003 is gratefully acknowledged.

Appendix

A.1 Public transport model

In order to compare the performance of the ODRP and the public transport systems, we now describe the public transport model we assume, following the classical model by Jansson (1980) and the posterior adaptations by Jara-Díaz and Gschwender (2009). We will describe in detail the circular model only, as the feeder one can be derived directly. Let us begin introducing some notation: T refers to the time required by a bus to tour the whole circuit, i.e.

$$T = \frac{Z \cdot a \cdot L}{v_1} \tag{A1}$$

Where L stands for the length of each arc. We assume that each user requires an average time t to board and alight the bus. Denoting by f the line frequency (to be optimised) and by Y the number of passengers per time unit, then the bus cycle time is:

$$t_c = T + \frac{tY}{f} \tag{A2}$$

To use Eq. (A2) to express the operators' costs, we use Eq. (10), i.e., assume that both operational and capital costs grow linearly with the bus capacity K. As the operating time is fixed in the public transport case (buses are operating all the time), this means that each bus cost can be expressed as $c_1 + c_2 K$, with $c_1 = c_{A1} + Ec_{O1}, c_2 = c_{A2} + Ec_{O2}$, where E is the total operation time. Operators' costs can then be written as:

$$f(T + tY/f)(c_1 + c_2K) \tag{A3}$$

Users' costs are a weighted average of waiting, walking, and in-vehicle times, through the respective parameters p_w,p_a, and p_v. Therefore, the public transport costs are calculated by solving the following optimization problem:

$$\min_{K,f} f(T + tY/f)(c_1 + c_2K) + Y(p_w t_w + p_a t_a + p_v t_v) \tag{A4}$$

$$\text{s.t. } K \geq \frac{t}{f} Y \tag{A5}$$

Eq. (A4) represents the sum of operators' and users' costs. We assume homogeneous headway, vehicles do not run full (passengers can board the first vehicle that arrives) and random user arrivals at constant rates, which imply that the average waiting time is $t_w = 1/2f$. Average in-vehicle time t_v can be calculated as we know the average distance travelled by the users; it includes running time plus time spent at stops where other users board and alight. Average walking distance can be computed directly when the random demand is created, by calculating the distances between the real origins and the bus stations of the respective zones, and doing the same for the destinations. Dividing such distances by the walking speed v_a results in the average walking time t_a. Eq. (A5) ensures that all users will fit on the bus. As the objective function in Eq. (A5) increases with K, this constraint will always be active. Factor α represents the ratio between the most loaded and the average arc, which can also be computed directly once the random demand is known.

A.2 Glossary and numerical value of the parameters

Table A1
Glossary of the parameters used throughout the paper. Stopping time τ is computed following Roess et al. (2004). Operators' cost parameters $c_{O1}, c_{O2}, c_{A1}, c_{A2}$ for human-driven and automated vehicles are calculated for Santiago, Chile, based on Tirachini and Antoniou (2020). Time required to board and alight the vehicles t is taken from Jara-Díaz et al. (2017). Walking speed v_A, as well as users' costs parameters p_w, p_a, p_v are obtained from Fielbaum et al. (2021). The rest of the parameters are ours.

Symbol	Meaning	Value
δ	Time elapsed between two consecutive assignments in ODRP.	1 [min]
τ	Time spent by the ODRP vehicle at each stop.	10.5 [sec]
a	Number of longitudinal streets in a zone.	5
b	Number of transversal streets in a zone.	7
v_1	Vehicles' speed in fast streets.	25 [km/h]
v_2	Vehicles' speed in low streets.	12.5 [km/h]
Z	Number of zones	45
γ	Level of dispersion of the origins and destinations within a zone.	0.2
l	Average number of zones toured by the users in the circular model.	10
σ^2	Variance of the number of zones toured by the users in the circular model.	4
L	Arcs' length.	50 [m]
t	Time required to board and alight a public transport vehicle.	5 [sec]
E	Total operation time	10 [h]
c_{O1}	Fixed operating cost per vehicle.	1.13 [US$/h]
c_{O2}	Capacity-dependant operating cost per vehicle.	0.074 [US$/h-seat]
c_{A1}	Fixed capital cost per vehicle (AV/Human-Driven).	24.6/78.1 [US$]
c_{A2}	Capacity-dependant capital cost per vehicle (AV/Human-Driven).	2.1/1.2 [US$/seat]
v_a	Walking speed.	5 [km/h]
p_v	Monetary equivalent cost of one time unit spent by a user in-vehicle.	2.32 [US$/h]
p_a	Monetary equivalent cost of one time unit spent by a user waiting.	4.64 [US$/h]
p_w	Monetary equivalent cost of one time unit spent by a user walking.	4.64 [US$/h]

References

Alonso-González, M.J., van Oort, N., Cats, O., Hoogendoorn-Lanser, S., Hoogendoorn, S., 2020. Value of time and reliability for urban pooled on-demand services. Transport. Res. C Emerg. Technol. 115, 102621.

Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E., Rus, D., 2017. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. Proc. Natl. Acad. Sci. USA 114 (3), 462–467. https://doi.org/10.1073/pnas.1611675114.

Arnott, R., 1996. Taxi travel should Be subsidized. J. Urban Econ. 40 (3), 316–333. https://doi.org/10.1006/juec.1996.0035.

Badia, H., Jenelius, E., 2021. Design and operation of feeder systems in the era of automated and electric buses. Transport. Res. Pol. Pract. 152, 146–172. https://doi.org/10.1016/j.tra.2021.07.015.

Bahrami, S., Nourinejad, M., Nesheli, M.M., Yin, Y., 2022. Optimal composition of solo and pool services for on-demand ride-hailing. Transport. Res. E Logist. Transport. Rev. 161, 102680 https://doi.org/10.1016/j.tre.2022.102680.

Basso, L.J., Jara-Díaz, S.R., 2006. Are returns to scale with variable network size adequate for transport industry structure analysis? Transport. Sci. 40 (3), 259–268. https://doi.org/10.1287/trsc.1060.0154.

Baumol, W.J., 1986. Contestable markets: an uprising in the theory of industry structure. Microtheory: Appl. Origins 40–54.

Bischoff, J., Führer, K., Maciejewski, M., 2019a. Impact assessment of autonomous DRT systems. Transport. Res. Procedia 41, 440–446. https://doi.org/10.1016/j.trpro.2019.09.074.

Bischoff, J., Führer, K., Maciejewski, M., 2019b. Impact assessment of autonomous DRT systems. Transport. Res. Procedia 41, 440–446. https://doi.org/10.1016/j.trpro.2019.09.074.

Bösch, P.M., Becker, F., Becker, H., Axhausen, K.W., 2018. Cost-based analysis of autonomous mobility services. Transport Pol. 64, 76–91. https://doi.org/10.1016/j.tranpol.2017.09.005.

Bürstlein, J., López, D., Farooq, B., 2021. Exploring first-mile on-demand transit solutions for North American suburbia: a case study of Markham, Canada. Transport. Res. Pol. Pract. 153, 261–283. https://doi.org/10.1016/j.tra.2021.08.018.

Calabrò, G., Araldo, A., Oh, S., Seshadri, R., Inturri, G., Ben-Akiva, M., 2021. Integrating fixed and demand-responsive transportation for flexible transit network design. In: TRB 2021: 100th Annual Meeting of the Transportation Research Board.

Cáp, M., Alonso-Mora, J., 2018. Multi-objective analysis of ridesharing in automated mobility-on-demand. In: *Proceedings Of RSS 2018: Robotics - Science And Systems XIV*. RSS 2018: Robotics - Science and Systems XIV. https://research.tudelft.nl/en/publications/multi-objective-analysis-of-ridesharing-in-automated-mobility-on-.

Castillo, J.C., Knoepfle, D., Weyl, G., 2017. Surge pricing solves the wild goose chase. In: Proceedings of the 2017 ACM Conference on Economics and Computation, pp. 241–242. https://doi.org/10.1145/3033274.3085098.

Chen, P., Nie, Y., 2017. Connecting e-hailing to mass transit platform: analysis of relative spatial position. Transport. Res. C Emerg. Technol. 77, 444–461. https://doi.org/10.1016/j.trc.2017.02.013.

Chen, S., Wang, H., Meng, Q., 2020. Solving the first-mile ridesharing problem using autonomous vehicles. Comput. Aided Civ. Infrastruct. Eng. 35 (1), 45–60.

Cortés, C.E., Jara-Díaz, S., Tirachini, A., 2011. Integrating short turning and deadheading in the optimization of transit services. Transport. Res. Pol. Pract. 45 (5), 419–434. https://doi.org/10.1016/j.tra.2011.02.002.
Daganzo, C.F., Ouyang, Y., 2019. A general model of demand-responsive transportation services: from taxi to ridesharing to dial-a-ride. Transp. Res. Part B Methodol. 126, 213–224. https://doi.org/10.1016/j.trb.2019.06.001.
Daganzo, C.F., Ouyang, Y., Yang, H., 2020. Analysis of ride-sharing with service time and detour guarantees. Transp. Res. Part B Methodol. 140, 130–150. https://doi.org/10.1016/j.trb.2020.07.005.
Delle Site, P., Filippi, F., 1998. Service optimization for bus corridors with short-turn strategies and variable vehicle size. Transport. Res. Pol. Pract. 32 (1), 19–38. https://doi.org/10.1016/S0965-8564(97)00016-5.
Diao, M., Kong, H., Zhao, J., 2021. Impacts of transportation network companies on urban mobility. Nat. Sustain. 1–7.
Evans, A.W., Morrison, A.D., 1997. Incorporating accident risk and disruption in economic models of public transport. J. Transport Econ. Pol. 31 (2), 117–146.
Fagnant, D.J., Kockelman, K.M., 2018a. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. Transportation 45 (1), 143–158. https://doi.org/10.1007/s11116-016-9729-z.
Fagnant, D.J., Kockelman, K.M., 2018b. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. Transportation 45 (1). https://doi.org/10.1007/s11116-016-9729-z. Article 1.
Fielbaum, A., 2020. Strategic public transport design using autonomous vehicles and other new technologies. Int. J. Intell. Transport. Syst. Res. 18 (2). https://trid.trb.org/view/1720905.
Fielbaum, A., 2021. Optimizing a vehicle's route in an on-demand ridesharing system in which users might walk. J. Intel. Transport. Syst. https://doi.org/10.1080/15472450.2021.1901225.
Fielbaum, A., Alonso-Mora, J., 2020. Unreliability in ridesharing systems: measuring changes in users' times due to new requests. Transport. Res. C Emerg. Technol. 121, 102831.
Fielbaum, A., Bai, X., Alonso-Mora, J., 2021. On-demand Ridesharing with Optimized Pick-Up and Drop-Off Walking Locations. Transportation Research Part C: Emerging Technologies, 103061.
Fielbaum, A., Jara-Diaz, S., Gschwender, A., 2017. A parametric description of cities for the normative analysis of transport systems. Network. Spatial Econ. 17 (2), 343–365. https://doi.org/10.1007/s11067-016-9329-7.
Fielbaum, A., Jara-Diaz, S., Gschwender, A., 2020a. Beyond the Mohring effect: scale economies induced by transit lines structures design. Econ. Transport. 22, 100163 https://doi.org/10.1016/j.ecotra.2020.100163.
Fielbaum, A., Jara-Diaz, S., Gschwender, A., 2020b. Lines spacing and scale economies in the strategic design of transit systems in a parametric city. Res. Transport. Econ., 100991 https://doi.org/10.1016/j.retrec.2020.100991.
Henao, A., Marshall, W.E., 2019. The impact of ride-hailing on vehicle miles traveled. Transportation 46 (6), 2173–2194.
Ho, C.Q., Hensher, D.A., Mulley, C., Wong, Y.Z., 2018. Potential uptake and willingness-to-pay for Mobility as a Service (MaaS): a stated choice study. Transport. Res. Pol. Pract. 117, 302–318.
Hörcher, D., Tirachini, A., 2021. A review of public transport economics. Econ. Transport. 25, 100196.
Jansson, J.O., 1979. Marginal cost pricing of scheduled transport services: a development and generalisation of Turvey and mohring's theory of optimal bus fares. J. Transport Econ. Pol. 13 (3), 268–294.
Jansson, J.O., 1980. A simple bus line model for optimisation of service frequency and bus size. J. Transport Econ. Pol. 53–80.
Jara-Diaz, S., 2007. Transport economic theory. https://trid.trb.org/view/1134779.
Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2017. Optimal fleet size, frequencies and vehicle capacities considering peak and off-peak periods in public transport. Transport. Res. Pol. Pract. 106, 65–74. https://doi.org/10.1016/j.tra.2017.09.005.
Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2020. Strategies for transit fleet design considering peak and off-peak periods using the single-line model. Transp. Res. Part B Methodol. 142, 1–18. https://doi.org/10.1016/j.trb.2020.09.012.
Jara-Díaz, S., Gschwender, A., 2003. Towards a general microeconomic model for the operation of public transport. Transport Rev. 23 (4), 453–469.
Jara-Díaz, S., Latournerie, A., Tirachini, A., Quitral, F., 2022. Optimal pricing and design of station-based bike-sharing systems: a microeconomic model. Econ. Transport. 31, 100273 https://doi.org/10.1016/j.ecotra.2022.100273.
Jara-Díaz, S.R., Gschwender, A., 2009. The effect of financial constraints on the optimal design of public transport services. Transportation 36 (1), 65–75.
Jara-Díaz, S., Tirachini, A., 2013. Urban bus transport: open all doors for boarding. J. Transport Econ. Pol.(JTEP) 47 (1), 91–106.
Kaddoura, I., Schlenther, T., 2021. The impact of trip density on the fleet size and pooling rate of ride-hailing services: a simulation study. Proc. Comput. Sci. 184, 674–679. https://doi.org/10.1016/j.procs.2021.03.084.
Kang, D., Levin, M.W., 2021. Maximum-stability dispatch policy for shared autonomous vehicles. Transp. Res. Part B Methodol. 148, 132–151. https://doi.org/10.1016/j.trb.2021.04.011.
Ke, J., Yang, H., Zheng, Z., 2020. On ride-pooling and traffic congestion. Transp. Res. Part B Methodol. 142, 213–231. https://doi.org/10.1016/j.trb.2020.10.003.
Kim, M.E., Schonfeld, P., 2014. Integration of conventional and flexible bus services with timed transfers. Transp. Res. Part B Methodol. 68, 76–97.
König, A., Grippenkoven, J., 2020. Travellers' willingness to share rides in autonomous mobility on demand systems depending on travel distance and detour. Travel Behav. Soc. 21, 188–202.
Lavieri, P.S., Bhat, C.R., 2019. Modeling individuals' willingness to share trips with strangers in an autonomous vehicle future. Transport. Res. Pol. Pract. 124, 242–261.
Lee, E., Cen, X., Lo, H.K., 2021. Zonal-based flexible bus service under elastic stochastic demand. Transport. Res. E Logist. Transport. Rev. 152, 102367 https://doi.org/10.1016/j.tre.2021.102367.
Leffler, D., Burghout, W., Jenelius, E., Cats, O., 2021. Simulation of fixed versus on-demand station-based feeder operations. Transport. Res. C Emerg. Technol. 132, 103401 https://doi.org/10.1016/j.trc.2021.103401.
Lehe, L., Gayah, V.V., Pandey, A., 2021. Increasing returns to scale in carpool matching: evidence from *scoop*. Findings 25093. https://doi.org/10.32866/001c.25093.
Levin, M.W., Kockelman, K.M., Boyles, S.D., Li, T., 2017. A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application. Comput. Environ. Urban Syst. 64, 373–383. https://doi.org/10.1016/j.compenvurbsys.2017.04.006.
Li, R., Qin, L., Yu, J.X., Mao, R., 2016. Optimal multi-meeting-point route search. IEEE Trans. Knowl. Data Eng. 28 (3), 770–784. https://doi.org/10.1109/TKDE.2015.2492554.
Li, W., Pu, Z., Li, Y., Tu, M., 2021. How does ride splitting reduce emissions from ridesourcing? A spatiotemporal analysis in Chengdu, China. Transport. Res. Transport Environ. 95, 102885 https://doi.org/10.1016/j.trd.2021.102885.
Li, X., Hu, S., Fan, W., Deng, K., 2018. Modeling an enhanced ridesharing system with meet points and time windows. PLoS One 13 (5), e0195927. https://doi.org/10.1371/journal.pone.0195927.
Li, X., Quadrifoglio, L., 2010. Feeder transit services: choosing between fixed and demand responsive policy. Transport. Res. C Emerg. Technol. 18 (5), 770–780. https://doi.org/10.1016/j.trc.2009.05.015.
Lokhandwala, M., Cai, H., 2018. Dynamic ride sharing using traditional taxis and shared autonomous taxis: a case study of NYC. Transport. Res. C Emerg. Technol. 97, 45–60. https://doi.org/10.1016/j.trc.2018.10.007.
Lotze, C., Marszal, P., Schröder, M., Timme, M., 2022. Dynamic stop pooling for flexible and sustainable ride sharing. New J. Phys. 24 (2), 023034 https://doi.org/10.1088/1367-2630/ac47c9.
Ma, T.-Y., Rasulkhani, S., Chow, J.Y.J., Klein, S., 2019. A dynamic ridesharing dispatch and idle vehicle repositioning strategy with integrated transit transfers. Transport. Res. E Logist. Transport. Rev. 128, 417–442. https://doi.org/10.1016/j.tre.2019.07.002.
Manik, D., Molkenthin, N., 2020. Topology dependence of on-demand ride-sharing. Appl. Network Sci. 5 (1) https://doi.org/10.1007/s41109-020-00290-2. Article 1.
Martinez, L.M., Viegas, J.M., 2017. Assessing the impacts of deploying a shared self-driving urban mobility system: an agent-based model applied to the city of Lisbon, Portugal. Int. J. Transport. Sci. Technol. 6 (1), 13–27. https://doi.org/10.1016/j.ijtst.2017.05.005.
Militão, A.M., Tirachini, A., 2021a. Optimal fleet size for a shared demand-responsive transport system with human-driven vs automated vehicles: a total cost minimization approach. Transport. Res. Pol. Pract. 151, 52–80. https://doi.org/10.1016/j.tra.2021.07.004.
Militão, A.M., Tirachini, A., 2021b. Optimal fleet size for a shared demand-responsive transport system with human-driven vs automated vehicles: a total cost minimization approach. Transport. Res. Pol. Pract. 151, 52–80. https://doi.org/10.1016/j.tra.2021.07.004.
Mohring, H., 1972. Optimization and scale economies in urban bus transportation. Am. Econ. Rev. 62 (4), 591–604.
Papanikolaou, A., Basbas, S., 2020. Analytical models for comparing Demand Responsive Transport with bus services in low demand interurban areas. Transport. Lett. 1–8.
Pimenta, V., Quilliot, A., Toussaint, H., Vigo, D., 2017. Models and algorithms for reliability-oriented Dial-a-Ride with autonomous electric vehicles. Eur. J. Oper. Res. 257 (2), 601–613. https://doi.org/10.1016/j.ejor.2016.07.037.
Pinto, H.K.R.F., Hyland, M.F., Mahmassani, H.S., Verbas, I.Ö., 2020. Joint design of multimodal transit networks and shared autonomous mobility fleets. Transport. Res. C Emerg. Technol. 113, 2–20. https://doi.org/10.1016/j.trc.2019.06.010.
Quadrifoglio, L., Li, X., 2009. A methodology to derive the critical demand density for designing and operating feeder transit services. Transp. Res. Part B Methodol. 43 (10), 922–935.
Roess, R.P., Prassas, E.S., McShane, W.R., 2004. Traffic Engineering. Pearson/Prentice Hall.
Roy, S., Cooper, D., Mucci, A., Sana, B., Chen, M., Castiglione, J., Erhardt, G.D., 2020. Why is traffic congestion getting worse? A decomposition of the contributors to growing congestion in San Francisco-Determining the Role of TNCs. Case Stud. Trans. Pol. 8 (4), 1371–1382. https://doi.org/10.1016/j.cstp.2020.09.008.
Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H., Ratti, C., 2014. Quantifying the benefits of vehicle pooling with shareability networks. In: Proceedings of the National Academy of Sciences, vol. 111, pp. 13290–13294. https://doi.org/10.1073/pnas.1403657111, 37.
Santos, D.O., Xavier, E.C., 2015. Taxi and ride sharing: a dynamic dial-a-ride problem with money as an incentive. Expert Syst. Appl. 42 (19), 6728–6737. https://doi.org/10.1016/j.eswa.2015.04.060.
Simonetto, A., Monteil, J., Gambella, C., 2019. Real-time city-scale ridesharing via linear assignment problems. Transport. Res. C Emerg. Technol. 101, 208–232. https://doi.org/10.1016/j.trc.2019.01.019.
Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., Pavone, M., 2014. Toward a Systematic Approach to the Design and Evaluation of Automated Mobility-On-Demand Systems: A Case Study in Singapore. En *Road vehicle Automation*. Springer, pp. 229–245.
Stiglic, M., Agatz, N., Savelsbergh, M., Gradisar, M., 2015. The benefits of meeting points in ride-sharing systems. Transp. Res. Part B Methodol. 82, 36–53. https://doi.org/10.1016/j.trb.2015.07.025.

Tachet, R., Sagarra, O., Santi, P., Resta, G., Szell, M., Strogatz, S.H., Ratti, C., 2017. Scaling law of urban ride sharing. Sci. Rep. 7 (1) https://doi.org/10.1038/srep42868. Article 1.

Tikoudis, I., Martinez, L., Farrow, K., García Bouyssou, C., Petrik, O., Oueslati, W., 2021. Ridesharing services and urban transport CO2 emissions: simulation-based evidence from 247 cities. Transport. Res. Transport Environ. 97, 102923 https://doi.org/10.1016/j.trd.2021.102923.

Tirachini, A., Antoniou, C., 2020. The economics of automated public transport: effects on operator cost, travel time, fare and subsidy. Econ. Transport. 21, 100151 https://doi.org/10.1016/j.ecotra.2019.100151.

Tirachini, A., Chaniotakis, E., Abouelela, M., Antoniou, C., 2020. The sustainability of shared mobility: can a platform for shared rides reduce motorized traffic in cities? Transport. Res. C Emerg. Technol. 117, 102707 https://doi.org/10.1016/j.trc.2020.102707.

Tirachini, A., Gomez-Lobo, A., 2020. Does ride-hailing increase or decrease vehicle kilometers traveled (VKT)? A simulation approach for Santiago de Chile. Int. J. Sustain. Transport. 14 (3), 187–204.

Tirachini, A., Hensher, D.A., 2011. Bus congestion, optimal infrastructure investment and the choice of a fare collection system in dedicated bus corridors. Transp. Res. Part B Methodol. 45 (5), 828–844. https://doi.org/10.1016/j.trb.2011.02.006.

Tirachini, A., Hensher, D.A., Jara-Díaz, S.R., 2010a. Comparing operator and users costs of light rail, heavy rail and bus rapid transit over a radial public transport network. Res. Transport. Econ. 29 (1), 231–242. https://doi.org/10.1016/j.retrec.2010.07.029.

Tirachini, A., Hensher, D.A., Jara-Díaz, S.R., 2010b. Restating modal investment priority with an improved model for public transport analysis. Transport. Res. E Logist. Transport. Rev. 46 (6), 1148–1168. https://doi.org/10.1016/j.tre.2010.01.008.

Tirachini, A., Hensher, D.A., Rose, J.M., 2013. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. Transport. Res. Pol. Pract. 53, 36–52. https://doi.org/10.1016/j.tra.2013.06.005.

Tsao, M., Milojevic, D., Ruch, C., Salazar, M., Frazzoli, E., Pavone, M., 2019. Model predictive control of ride-sharing autonomous mobility-on-demand systems. In: *2019 International Conference On Robotics And Automation* (ICRA), pp. 6665–6671. https://doi.org/10.1109/ICRA.2019.8794194.

Turvey, R., Mohring, H., 1975. Optimal bus fares. J. Transport Econ. Pol. 9 (3), 280–286.

van Engelen, M., Cats, O., Post, H., Aardal, K., 2018. Enhancing flexible transport services with demand-anticipatory insertion heuristics. Transport. Res. E Logist. Transport. Rev. 110, 110–121.

van Lierop, D., Badami, M.G., El-Geneidy, A.M., 2018. What influences satisfaction and loyalty in public transport? A review of the literature. Transport Rev. 38 (1), 52–72. https://doi.org/10.1080/01441647.2017.1298683.

Vazifeh, M.M., Santi, P., Resta, G., Strogatz, S.H., Ratti, C., 2018. Addressing the minimum fleet problem in on-demand urban mobility. Nature 557 (7706), 534–538.

Viergutz, K., Schmidt, C., 2019. Demand responsive - vs. conventional public transportation: a MATSim study about the rural town of Colditz, Germany. Proc. Comput. Sci. 151, 69–76. https://doi.org/10.1016/j.procs.2019.04.013.

Wallar, A., Van Der Zee, M., Alonso-Mora, J., Rus, D., 2018. Vehicle rebalancing for mobility-on-demand systems with ride-sharing. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4539–4546. https://doi.org/10.1109/IROS.2018.8593743.

Wang, Y., Zheng, B., Lim, E.-P., 2018. Understanding the effects of taxi ride-sharing—a case study of Singapore. Comput. Environ. Urban Syst. 69, 124–132. https://doi.org/10.1016/j.compenvurbsys.2018.01.006.

Wang, Z., Hyland, M.F., Bahk, Y., Sarma, N.J., 2022. On optimizing shared-ride mobility services with walking legs. https://arxiv.org/abs/2201.12639v1.

Ward, J.W., Michalek, J.J., Samaras, C., Azevedo, I.L., Henao, A., Rames, C., Wenzel, T., 2021. The impact of Uber and Lyft on vehicle ownership, fuel economy, and transit across U.S. cities. iScience 24 (1), 101933. https://doi.org/10.1016/j.isci.2020.101933.

Wen, J., Chen, Y.X., Nassir, N., Zhao, J., 2018. Transit-oriented autonomous vehicle operation with integrated demand-supply interaction. Transport. Res. C Emerg. Technol. 97, 216–234.

Wu, X., MacKenzie, D., 2021. Assessing the VMT effect of ridesourcing services in the US. Transport. Res. Transport Environ. 94, 102816 https://doi.org/10.1016/j.trd.2021.102816.

Yan, C., Zhu, H., Korolko, N., Woodard, D., 2020. Dynamic pricing and matching in ride-hailing platforms. Nav. Res. Logist. 67 (8), 705–724. https://doi.org/10.1002/nav.21872.

Yu, X., Shen, S., 2020. An integrated decomposition and approximate dynamic programming approach for on-demand ride pooling. IEEE Trans. Intell. Transport. Syst. 21 (9), 3811–3820. https://doi.org/10.1109/TITS.2019.2934423.

Zhang, K., Nie, Y., 2021. To pool or not to pool: equilibrium, pricing and regulation. Transp. Res. Part B Methodol. 151, 59–90. https://doi.org/10.1016/j.trb.2021.07.001.