Diseconomies of scale and subsidies in urban public transportation

Nicolas Coulombel, Guillaume Monchambert

Version of Record: https://www.sciencedirect.com/science/article/pii/S0047272723000853
Manuscript_6a5d88a0dd6b0528fd183bfb06d61e29

Diseconomies of scale and subsidies in urban public transportation

Nicolas Coulombel [a,*] *and Guillaume Monchambert* [b]

[a] LVMT, UMR-T 9403, Ecole des Ponts, Université Gustave Eiffel, Champs-sur-Marne, France.

[*] Corresponding author: nicolas.coulombel@enpc.fr

[b] LAET, University of Lyon, Université Lyon 2, Lyon F69007, France.

Abstract

Subsidization of urban public transportation systems is often motivated by economies of scale and second-best considerations such as an underpriced road alternative. We model a public transit line subject to frictions between users (in-vehicle crowding), users and vehicles (boarding and alighting delays), and vehicles (congestion). We derive the profit- and welfare-maximizing provisions of supply. We show that if demand exceeds a first threshold, the system enters a congested regime and service frequency decreases. Beyond a second threshold, the strong deterioration of service quality causes the transit line to operate under diseconomies of scale, calling for a Pigouvian tax instead of a subsidy. This finding, which goes against Mohring's classical rule (1972), holds with an untolled road alternative, provided that the network structure remains constant. We estimate the model for the London Piccadilly line and find evidence of substantial diseconomies of scale during the morning peak, adding up to -1.49 £/trip for the observed provision of service quality (-0.61 £/trip at optimum). These results question current subsidy policies for the busiest transit lines.

JEL Codes: D42; D62; H24; R41; R48

Keywords: congestion; mass transit; externality; Mohring effect; London Piccadilly line

1. Introduction

Public transport accounts for a large share of travel in major cities: around 47% in Tokyo, Japan, 35% in London, UK, 27 % in Sydney, Australia, and 24% in Toronto, Canada (Deloitte City Mobility Index 2020). Despite or explaining this, urban public transportation systems are heavily subsidized (Table 1). The high levels of subsidies are a subject of increasing concern. First, they undermine the investment capacity of public transit authorities, while the development of public transit - along with active modes and shared mobility - is a key component of transport decarbonization strategies in virtually all cities. Moreover, low public transit fares generate high demand which can lead to excessive congestion, reducing the efficiency and attractiveness of public transit systems.

Table 1: Farebox recovery ratios (ratio of fare revenue to operating costs) for public transportation systems

Country	City	Public Transit Authority	Farebox ratio (%)	Year
Hong Kong	Hong Kong	MTR	172	2018
Japan	Tokyo	Tokyo Metro	129	2018
USA	San Francisco	BART	83	2017
Singapore	Singapore	SMRT	75	2016
UK	London	TfL	64	2018-2019
Canada	Toronto	TTC	61	2018
USA	New York City	MTA	52	2018
France	Paris	IdF-M (formerly STIF)	48	2015
Belgium	Brussels	MRBC	47	2018
Australia	Sydney	TfNSW	22	2017-2018
USA	Los Angeles	LACMTA	17	2018

Note: Figures have been computed by the authors or retrieved from the following sources: MTR Annual report 2018, p.211 (Hong Kong), Tokyo Metro Corporate Profile 2019, p.33 (Tokyo), San Francisco Bay Area Rapid Transit District Budget Summary Fiscal year 2018, p.5 (San Francisco), SMRT Corporation Ltd Annual Report 2016, p.34 and 35 (Singapore), TfL Annual Report and Statement of Accounts 2018/19, p.128 and 129 (London), 2018 Annual Report Toronto Transit Commission, p.17 and 43 (Toronto), Metropolitan Transportation Authority Financial Statements for the Years Ended December 31, 2018 and 2017, p.12 (New-York City), Activity Report 2015 STIF, p.7 and 8 (Paris), Statistics 2018 STIB, p.3 (Brussels), Transport 2018, p.10 (Sydney), and Comprehensive Annual Financial Report For the Fiscal Year Ended June 30, 2018, p.160 (Los Angeles).

The economic literature advances two main rationales for subsidizing public transport (Parry and Small, 2009). First, public transit systems operate under economies of scale. These arise from production costs (Farsi et al., 2007; Ripplinger and Bitzan, 2018; Viton, 1992), but also from user costs, as an increase in demand leads the operator to improve service quality, causing the average user cost to fall (Mohring, 1972). Second, car travel is typically underpriced relatively to the external costs that it generates, partly due to the political difficulty of introducing road pricing (De Borger and Proost, 2012). As such, subsidizing public transportation represents a second-best solution to support modal shift and to limit car use (Adler and van Ommeren, 2016; Anderson, 2014; Glaister and Lewis,

1978; Nelson et al., 2007; Parry, 2002).[1] While the relevance of each rationale may vary depending on the city characteristics, empirical studies have typically found substantial economies of scale related to user costs (Nash et al., 2001; Nelson et al., 2007; Parry and Small, 2009; Savage, 2010), justifying the central role of the corresponding "Mohring effect" in the literature.

This paper investigates the effect of congestion on economies of scale in public transportation, including implications in terms of pricing and subsidies. Additional passengers degrade service quality because they delay trains in the station by boarding and alighting, and increase crowding. In principle, the operator could compensate by increasing frequency or vehicle size; yet a minimum safe headway between trains puts an upper limit on service frequency, which compounds the previous two effects. This causes service quality to strongly deteriorate as demand becomes too high, thereby mitigating and in some cases even resulting in findings opposite to the previous literature, which did not consider these frequency constraints for public transit lines.

The analysis focuses on economies of density, defined here as economies of scale under constant network structure and transportation technology. We develop a model that captures several key features of public transit congestion (in-vehicle crowding, effects on dwelling time and frequency), allowing to investigate the effect of increasing levels of demand on service quality (frequency, vehicle capacity) and (dis)economies of density. Our model builds on the theoretical framework developed by Mohring (1972).[2] This framework has been widely applied to study public transit operations, economies of scale and optimal pricing (e.g. Basso and Silva, 2014). Yet, a key assumption underlying the Mohring effect is that frequency increases with demand (Jara-Díaz and Gschwender, 2003a). In current urban context, public transit ridership has increased to such an extent that this assumption does not hold anymore: there are increasing cases of heavily congested lines for which frequency falls if demand is too strong, as a result of too many users seeking to board and alight at each station. Moreover, most empirical studies on the topic largely ignore crowding costs, which are nevertheless a crucial consumption externality characterizing urban public transportation (de Palma et al., 2017). Therefore, we include three types of frictions. Following Kraus (1991) and de Palma et al. (2017), frictions between users are modeled as a crowding cost, which increases with in-vehicle occupancy. Frictions between users and vehicles are represented by considering that the dwelling time varies according to the flow of boarding and alighting users, as in Turvey and Mohring (1975). Finally, frictions between vehicles are considered through a minimum safe headway between two successive vehicles. This constraint imposes a hard physical limit on service frequency.

We find that urban public transit operations are characterized by economies of density only up to a certain level of demand. If demand is too strong, the severity of congestion causes the marginal social cost of an extra passenger to soar and exceed the average social cost, implying diseconomies of scale. In the short run (fixed frequency and vehicle size), frictions between users (crowding) and between users and vehicles (variable boarding and alighting time) primarily account for diseconomies

[1] Other rationales include addressing the special needs for transit of the underprivileged who would be either unable (due to disabilities) or could not afford to drive or access other forms of transportation (Vickrey, 1980), or supporting efficient investment in the transportation system in the long run (Brueckner and Selod, 2006).

[2] See Jara-Díaz and Gschwender (2003a) for a review of Mohring's model and its various extensions.

of density. In the medium run (adjustable frequency only) and long run (adjustable frequency and vehicle size), we show under general assumptions frictions between vehicles to be the main source of diseconomies of density. As a corollary, the optimal subsidy becomes negative, implying the standard Pigouvian tax. Our findings are robust in presence of an unpriced substitute transportation mode (typically the car), meaning that second-best pricing does not necessarily make a case for subsidizing public transit users. An application to the Piccadilly line in London provides empirical evidence of substantial diseconomies of scale during the morning peak, both for the observed (-1.49 £/trip) and optimal (-0.61 £/trip) provisions of service quality. Failing to account for congestion between vehicles leads to greatly overestimating the Mohring effect, thus economies of density (with a true value of 0.15 £/trip instead of -0.61£/trip at optimum). In the off-peak period, the prevalence of the Mohring effect is reasserted, as lower crowding levels lead to normal operations and the usual economies of scale (0.28 £/trip for the observed situation and 0.22 £/trip at optimum). Last, we simulate the effects of the New Tube for London (NTfL) scheme, which will renew the signaling system and rolling stock of the Piccadilly line in order to improve service quality. This illustrates that technological improvements can help alleviating congestion and curbing diseconomies of scale (from -2.29 £/trip to -1.56 £/trip at optimum).

This paper contributes to the literature on public transportation congestion (recently reviewed in Zhang et al., 2019) by showing how (severe) congestion can lead to diseconomies of scale, in contrast to previous works which find economies of scale to decrease yet to persist when congestion increases. We prove under general assumptions that among the three frictions, only between-vehicle congestion can result in diseconomies of density regarding user costs if the operator is able to adjust frequency.[3] This study is also to the best of our knowledge the first to provide empirical evidence of diseconomies of density in urban public transportation.

The analysis focuses on the economics of congestion for busy transit lines with high frequencies. Accordingly, we consider non-planning users only.[4] To limit model complexity, we also do not explicitly represent the effect of congestion on waiting times through capacity constraints and denied boarding. Denied boarding being inefficient - because the operator will need to supply the capacity necessary to serve the denied users eventually -, it is a phenomenon typically related to frequency constraints.[5] Considering this effect would thus strengthen our key finding that among the three types of frictions, only frictions between vehicles can and do lead to diseconomies of scale under adjustable frequency. The economies of scale analyzed in this paper correspond to economies of

[3] We generalize previous findings from Hörcher (2017) and Tirachini et al. (2010a) who found diseconomies of scale in more restrictive contexts (hard constraints on frequency for the former, fixed vehicle size for the latter).

[4] This case corresponds to the situation where service frequency is sufficiently high so that users find it less costly to just go to the station and wait, rather than accessing the exact timetable information and synchronizing departure from home with the public transit schedule (Fosgerau, 2009; Jansson, 1993).

[5] First studied by Oldfield and Bly (1988), capacity constraints and denied boarding were further investigated by Kraus and Yoshida (2002) and Yoshida (2008). To do so, they use a dynamic model - the bottleneck model for Yoshida (2008) - instead of the steady state model of Mohring (1972), at the cost of greater analytical complexity. Kraus and Yoshida find that queueing can occur at equilibrium because of the finite fleet size and corresponding headway constraints, but that it disappears under optimal pricing. Denied boarding has also attracted attention in transit assignment modeling (see e.g. Cominetti and Correa, 2001).

density, linked to variations in travel demand under constant network structure and transportation technology. In the longer run, a sustained increase in demand can be met with two main decisions other than adjusting frequency or vehicle size: 1) improving the transportation technology, through the use of heavier and faster modes (Tirachini et al., 2010b), more efficient boarding/alighting technologies (Jara-Díaz and Tirachini, 2013), or better signaling systems, and 2) changing the network design (Badia et al., 2014; Daganzo, 2010) and line density (Chang and Schonfeld, 1991). As these levers allow to accommodate more demand and leverage more economies of scale (e.g. Fielbaum et al., 2020; Jara-Díaz and Gschwender, 2003b), they are expected to mitigate (technological improvements) or even offset (network improvements) diseconomies of density linked to between-vehicle congestion. This comes at the cost of substantial investments, as illustrated by the NTfL scheme. Considering funding constraints, diseconomies of density are therefore likely to remain a cause of concern in many transit networks around the world.

Finally, our results contrast with the standard transport economics literature in which public transport operates under economies of scale that justify subsidies (Mohring, 1972). On the contrary, the diseconomies of scale arising with high levels of demand call for a review of subsidies schemes in congested urban transportation systems and provide additional support for fare differentiation and peak pricing. These findings are of interest to the regulator and operators, but they may also lead to new results in urban and transport economics.

2. Motivating facts

This section details empirical facts relative to the effects of demand on public transit operations, with a focus on urban heavy rail systems (trains, subways), that motivate our study. The choice of cities studied here is both limited by the availability of freely accessible data and motivated by external validity, with the objective of showing that mass transit congestion phenomena are observable in networks in different countries across continents. We represent here situations of equilibrium between supply and demand, which are the consequences of the optimizing decisions of the agents who provide (operator and transport authority) and of the agents who use the service (passengers).

2.1. Service frequency

The effect of demand on frequency is non-monotonic. At first, transit authorities operate more trains to serve additional users, as demonstrated theoretically by Mohring (1972) and confirmed empirically (Parry and Small, 2009). As demand keeps growing further, the frequency eventually meets the physical limit of the network, however. The dwelling time needed for all passengers to board and alight increases to such a point that it causes the maximum feasible frequency to decrease (Canavan et al., 2019). This non-monotonic relationship between demand and frequency is exemplified for two very busy lines, Linea D in Buenos Aires, Argentina, and Piccadilly line in London, UK (Figure 1).

The decrease in frequency stems from the combination of increased boarding and alighting times and of the incompressibility of the safe headway, sometimes generating queues of trains on the tracks (Lam et al., 1998). This interplay between crowding and frequency has been observed in

most crowded metro systems over the world (see for example newspaper articles by Daozu, 2010 for Shanghai, China, by Fitzsimmons et al., 2017 and by Baker, 2018, for Tokyo, Japan).

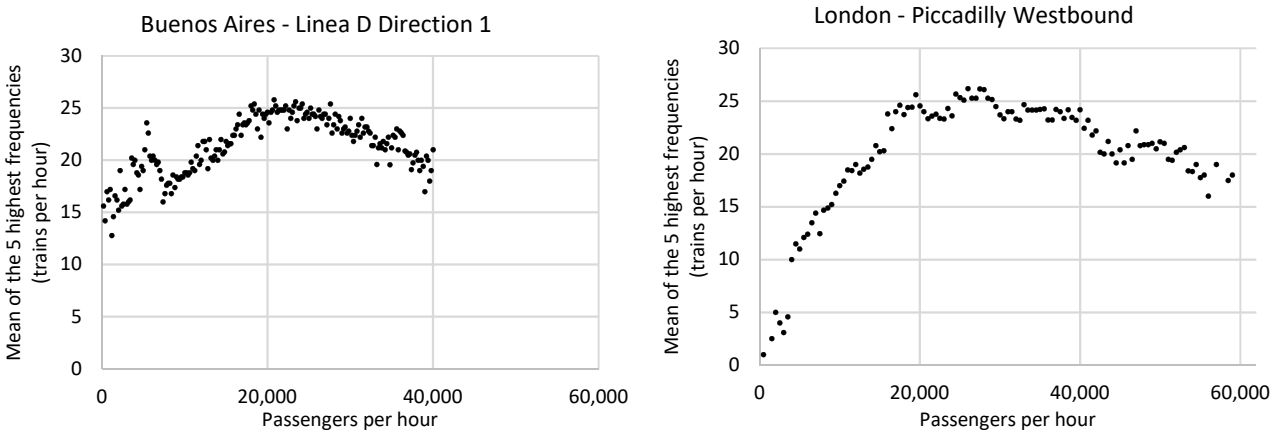


Figure 1: Maximum operated frequency as a function of demand.

Notes: The figures above show the maximum operated frequencies as a function of demand for Linea D direction 1 in Buenos Aires, Argentina (years 2015 to 2017), and Piccadilly line westbound, in London, UK (years 2013 and 2014). Each dot represents the mean of the five highest observed frequencies, measured in trains per hour, for given demand levels, measured in passengers per hour. Passenger counts come from ticketing data. If a station serves several lines, the data is weighted by the number of lines.

Sources: Subterráneos Buenos Aires (frequencies: Subte Trenes despachados; passengers: Subte Viajes Molinetes) ; Transport for London (frequencies: Freedom of Information request 1583-1819; passengers: Freedom of Information request 1276-1516).

2.2. Travel time

Another consequence of the effect of demand on dwelling times is longer in-vehicle travel times. This effect is amplified by congestion between trains. This happens when a train must halt and wait for the previous one to leave the station, as it can happen in the case of bus bunching (Daganzo, 2009). Again, we illustrate the effect of demand on travel times for two busy lines: the Metro Red line in Washington DC, US, and the RER B heavy rail line in Paris, France (Figure 2). As expected, travel times are longer during the morning and evening peaks when demand is stronger (grey areas).

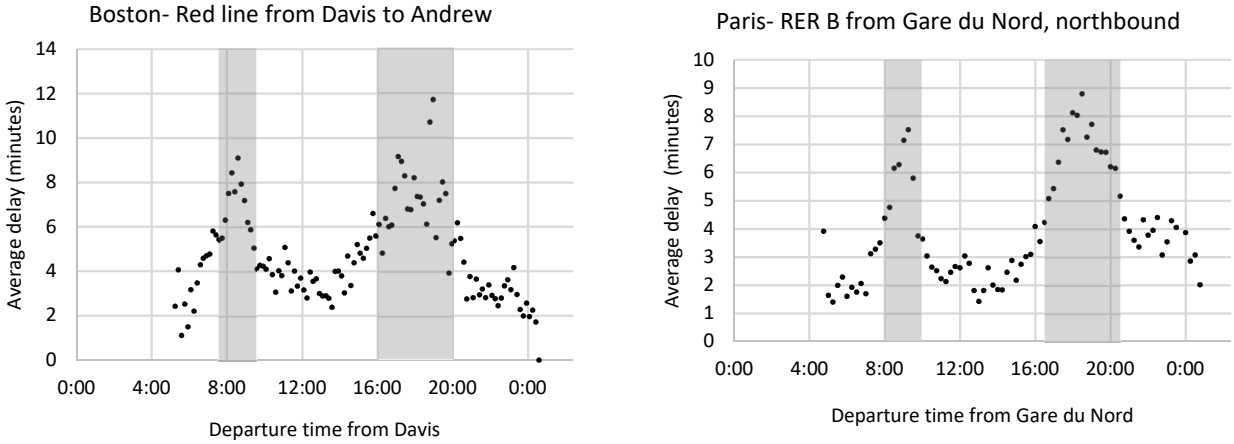


Figure 2 : Average delay as a function of departure time.

Notes: The figures above show the average delay (measured in minutes) for a trip from Davis to Andrew on red line in Boston, US, and for a trip from Gare du Nord to the north on RER B in Paris, France. The data was

collected on working days, from 2nd to 13th December 2019 for Boston and from 1st July to 31st December 2019 for Paris. Peak periods are highlighted with grey area.
Sources: Boston travel times: MBTA Rapid Transit Travel Times 2019 (Massachusetts Bay Transportation Authority); Paris travel times: Détails des circulations quotidiennes des trains SNCF d'Île-de-France.

2.3. Occupancy rate

The occupancy rate is the ratio between travel demand (in passenger-kilometers) and transportation supply (in vehicle-kilometers or seat-kilometers). This ratio would remain constant were supply to vary commensurately with demand along the day. However, several works have shown that supply should generally react less than linearly to demand variations - both intra-day (peak/off-peak imbalances) and inter-day (evolutions over time) – in order to reduce system costs (Jansson, 1993; Mohring, 1972).[6] Empirically, demand does increase faster than supply in peak periods, leading to greater occupancy rates and to a decrease in comfort and in the perceived quality of the trip. This relationship between occupancy rates (proxied by the average number of passengers per train) and demand (proxied by time of day) is illustrated for the Linea D in Buenos Aires, Argentina and the Underground Piccadilly line in London, England (Figure 3).

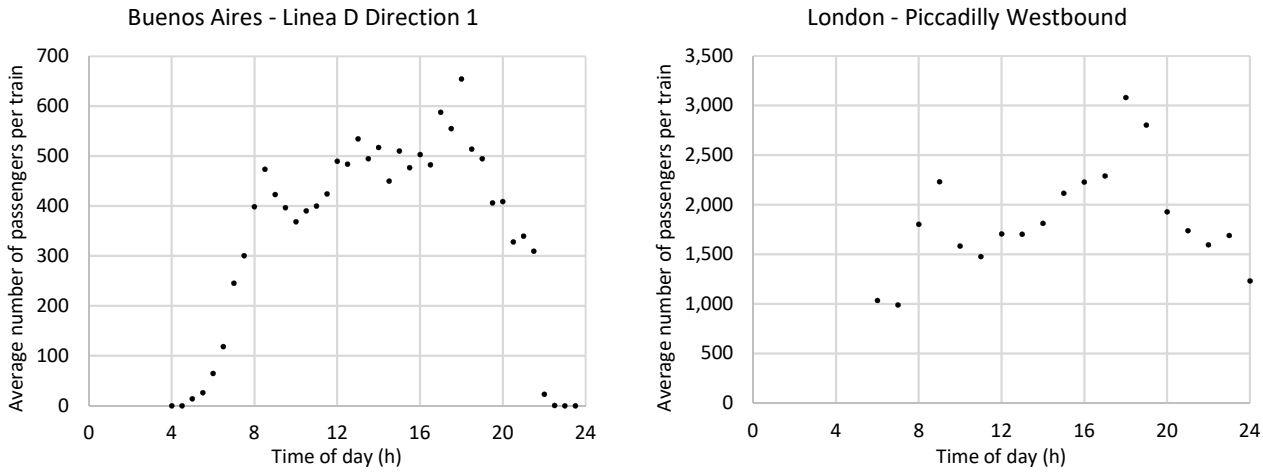


Figure 3: Average number of passengers per train as a function of departure time.
Notes: The figures above show the average number of passengers, measured from ticketing data, per train, measured from frequency per hour, for Linea D direction 1 in Buenos Aires, Argentina (years 2015 to 2017), and Piccadilly line westbound, in London, UK (years 2013 and 2014). If a station serves several lines, the data is weighted by the number of lines. Dots are not occupancy rate as distances travelled by passengers is unknown.
Sources: Buenos Aires frequencies: Subte Trenes despachados (Subterráneos Buenos Aires); Buenos Aires passengers: Subte Viajes Molinetes (Subterráneos Buenos Aires); London frequencies: Freedom of Information request 1583-1819 (Transport for London); London passengers: Freedom of Information request 1276-1516 (Transport for London).

[6] There are some exceptions to this general rule, e.g. if the off-peak conditions are such that vehicles are full during that period, then the optimal frequency does increase linearly with daily demand (Jara-Díaz et al., 2020).

3. Model

Building on empirical evidence, we extend the transit line model of Mohring (1972) by including the effects described above, and analyze its new properties. Without loss of generality, we will assume the transit line to be a railway line and refer to it as such in the remainder of the paper.

This section focuses on the general case. For concision, proofs and detailed computations (including a notational glossary) are sent back to Appendix. The special case with linear specifications, which we use to illustrate our results, is developed as an online appendix.

3.1. Set-up

Consider a mass transit line with stations evenly spaced by a distance d_S. We study the steady state of a one kilometer route segment over a given time period, typically one hour during the morning peak. In each station, new users arrive at a constant rate N (in users/h/km of route). [7] For model tractability, trip length is assumed constant and equal to *d*.

3.1.1. Transportation technology

Let *F* denote the service frequency and $H \equiv F^{-1}$ the headway. The service is regular and reliable, meaning that *H* is constant over the time period considered.[8] Vehicle size (capacity) is denoted by s.

From the model assumptions, the numbers of users alighting (n_A) and boarding (n_B) are the same at each station and for each train: $n_A = n_B = d_S\, N/F$. Passengers stay onboard for d/d_S stations, hence a vehicle load equal to dN/F. The load factor is then defined as the vehicle load over capacity:

$$l = \frac{dN}{sF}. \tag{1}$$

The total travel time is the sum of access time, waiting time and in-vehicle travel time. As interstation distance (and line density) remains constant throughout the analysis, access time is assumed to be null without loss of generality.

Because headways are regular and users arrive at a constant rate, the average waiting time is half the headway:

$$t_W = \frac{1}{2F}. \tag{2}$$

In-vehicle travel time is given by:

$$t_V = \frac{d}{v} + \frac{d}{d_S}\delta\left(\frac{N}{F}\right). \tag{3}$$

[7] The assumption of non-planning users who do not look at the schedule and therefore arrive at a constant rate over time is consistent with our focus on public transit congestion under high frequencies and levels of demand. See Tirachini et al. (2010a) for an extension of Mohring's framework to low headways and planning users.

[8] Following Benezech and Coulombel (2013), the model could be extended to the case of variable headways. This would involve substantially greater analytical complexity unless assuming linear utility functions, however.

The term d/v is the fixed free-flow travel time, where $v = (1/v_0 + \delta_0/d_S)^{-1}$ depends on the cruising speed between two stops v_0 and on the fixed dwelling time per stop δ_0. The term $d/d_S\delta(N/F)$ is the total variable dwelling time. The extra dwelling time per stop $\delta(N/F)$ is the sum of the alighting time $\delta_A(n_A/n_D)$ and boarding time $\delta_B(n_B/n_D)$, which depend on the ratios between the volumes of users alighting and boarding and the number of doors n_D. This yields $\delta(x) = \delta_A(d_S x/n_D) + \delta_B(d_S x/n_D)$.[9] The function δ is further assumed to be a strictly increasing and non-bounded C^1 function. Accordingly, in-vehicle travel time strictly increases with the level of demand *N* while it strictly decreases with *F* as greater frequencies allow for shorter boarding-alighting times.[10]

In Mohring's original model, optimal frequency tends toward infinity as demand keeps increasing. Yet, service frequency is subject to operational constraints. First, the headway cannot be lower than the dwelling time. Moreover, regulators enforce an additional minimum safe headway $H_0 > 0$ to limit collisions. This implies the following constraint on headways: $H \geq H_0 + \delta_0 + \delta(NH)$. Let $\bar{H}(N)$ denote the first positive solution to $H = H_0 + \delta_0 + \delta(NH)$ and $\bar{F}(N) = \bar{H}(N)^{-1}$. Rewriting the headway constraint in terms of frequency leads to:

$$F \leq \bar{F}(N), \tag{4}$$

where the maximum feasible frequency $\bar{F}(N)$ decreases with *N* (Appendix C). As demand increases, more time is required for allowing passengers to alight and to board, fewer trains can pass and the maximum frequency decreases. We refer to "overcrowding" as the situation in which demand is such that it is no longer possible to raise frequency, *i.e.* constraint (4) is active. For a planned frequency F, overcrowding occurs when $N \geq \bar{F}^{-1}(F)$, causing the real frequency $\bar{F}(N)$ to be lower than F.

3.1.2. Production costs

Transit operations imply production costs, which are assumed to be ultimately supported by the transit agency.[11] As the model represents a single line and does not account for variable line density, we overlook infrastructure costs (as in Parry and Small, 2009). The operating cost per kilometer of route C_{TA} includes vehicle capital costs and other operating costs, which depend on two main outputs, vehicle-kilometers X and vehicle-hours Z, and on vehicle size s: $C_{TA}(X, Z, s) = C_K(sX) + C_0(Z)$. Capital cost $C_K(sX)$ captures the depreciation of vehicles, which is assumed to depend on the product of vehicle-kilometers and vehicle size (i.e. on seat-kilometers supplied). Other operating

[9] The reduced form $\delta(N/F)$ implicitly treats the number of doors as fixed and independent from vehicle size *s*. This corresponds to the situation where the number of cars per train is fixed, typically due to length constraints (as the train length may not exceed that of the platform). Capacity is adjusted either by rearranging the interior of the cars or by expanding the size of each car while keeping the number of openings constant (such as by switching from single-decker to double-decker trains). The opposite polar case where the number of doors increases linearly with vehicle size *s*, leading to a reduced form $\delta(N/sF)$, is considered in the sensitivity analysis.

[10] For simplicity of exposition, we assume in the model that the running speed of trains is fixed and independent of passenger load: the in-vehicle travel time externality only results from the marginal boarding/alighting time. However, the $\delta(N/F)$ could in fact also include the effect of passenger load (which is commensurate with N/F) on running speed, so that there is no loss of generality.

[11] We ignore the potential contracts issues between the transit authority and the transport operator. These issues have been studied by Gagnepain and Ivaldi (2002), among others.

costs $C_0(Z)$ are based on vehicle-hours, and correspond to the cost of drivers as well as other time-based costs.[12] We assume that the cost functions C_K and C_0 are strictly increasing C^2 functions.

At the steady state, vehicle-kilometers (per kilometer of route) are given by $X = F$. To operate one kilometer of line with frequency F, the required number of vehicles is the ratio between the train runtime and the headway (Kraus and Yoshida, 2002), which yields: $Z = F \cdot (1/v + \delta(N/F)/d_S)$. Productions costs can therefore be rewritten as a function of frequency, vehicle size, and demand:

$$C_{TA}(F, s, N) = C_K(sF) + C_0\left(\frac{F}{v} + \frac{F}{d_S}\delta\left(\frac{N}{F}\right)\right). \quad (5)$$

3.1.3. Demand

Demand is characterized by the generalized inverse demand function $G(N)$.[13] The monetary reservation price is $P(N) = G(N) - C_U$, where one subtracts the user generalized travel cost C_U from $G(N)$.

The user cost C_U is a function of waiting time, in-vehicle travel time, and the level of crowding. To keep the model tractable, we assume that it is additively separable in each of its cost components, which are strictly increasing and convex C^2 functions: $C_U(t_W, t_V, l) = C_W(t_W) + C_V(t_V) + C_C(l)$. [14] From (1), (2) and (3), C_U rewrites as a function of frequency, vehicle size and demand as follows:[15]

$$C_U(F, s, N) = C_W\left(\frac{1}{F}\right) + C_V\left(\delta\left(\frac{N}{F}\right)\right) + C_C\left(\frac{dN}{sF}\right). \quad (6)$$

User costs are subject to two sources of externality: an additional user increases the crowding cost, but also the in-vehicle cost (through longer dwelling times).[16]

[12] The term $C_K(sX)$ may also account for distance-based operating costs (e.g. fuel consumption) but for simplicity we refer to this term as "capital costs".

[13] Here we assume that the inverse demand function is independent of service quality. See Basso and Jara-Diaz (2010) for a discussion of this assumption and under which conditions it holds.

[14] This implies in particular that the crowding cost $C_C(l)$ is independent of in-vehicle travel time t_V, as empirically supported by the study of De Lapparent and Koning (2016), among others. Assuming the crowding cost to vary with in-vehicle travel time makes the algebra substantially more complex, while not changing the main results (as illustrated in the sensitivity analysis). The waiting cost function - in a broad definition which encompasses scheduling costs and information costs relative to the timetable for low frequencies - is assumed to be convex, as demonstrated by Fosgerau (2009) under general conditions.

[15] Here we proceed to slight changes regarding the definitions of C_W and C_V by omitting some constant terms to simplify notations. We do not rename C_W and C_V to avoid notation clutter.

[16] Congestion inside the station (on platforms, escalators or walkways) and failures to board, which imply longer waiting times while adding to platform congestion, are not explicitly included in the model. Platform congestion depends on the stock of candidates for boarding, which between two trains equals on average $n_B/2 = d_S\, N/2F$. Because it varies like *N/F*, the cost could be integrated within the general in-vehicle cost function $C_V\big(\delta(N/F)\big)$. Similarly, the probability to fail to board being linked to the vehicle occupancy rate dN/sF, the corresponding cost could be integrated within the crowding cost function by allowing it to rise beyond a 100% occupancy rate. This suggests that our main findings are robust to the inclusion of other forms of congestion.

3.1.4. Model closure

We consider three provision regimes: welfare-maximizing, profit-maximizing, and welfare-maximizing with a marginal cost of public funds (MCPF), denoted by superscript $*$, e and μ respectively.

Consider first that the transit authority maximizes social welfare:

$$SW(F,s,N) = \int_0^N G(n)dn - SC(F,s,N). \tag{7}$$

where $SC(F,s,N) = N.C_U(F,s,N) + C_{TA}(F,s,N)$ denotes the system social cost (per kilometer of steady state route and per hour). As the aggregate gross user benefit $\int_0^N G(n)dn$ is independent of *F* and *s*, welfare maximization leads to a bi-level optimization problem: 1) for a given demand level *N*, choosing *F* and *s* so as to minimize the social cost, and 2) optimal choice of *N* at the upper level.

The monopoly and MCPF regimes lead to the same provision rules for service quality than at optimum (see Appendix): $s^*(N) = s^\mu(N) = s^e(N)$ and $F^*(N) = F^\mu(N) = F^e(N)$.[17] Because the level of demand varies across regimes (with $N^e \leq N^\mu \leq N^*$) , so does the level of service quality, however. Considering the similarities, we will focus the analysis and mathematical proofs on the optimum.

Based on the previous considerations, we first discuss the optimal provision of service quality ($F^*(N)$ and $s^*(N)$) at the lower level, then optimal pricing rules (and demand) at the upper level.

3.2. Results

3.2.1. Optimal service quality

Given the level of demand N, the transit authority supplies service frequency F and vehicle size s so as to minimize the social cost, subject to the frequency constraint:

$$\min_{F,s} SC(F,s,N) \quad s.t.\ F \leq \bar{F}(N) \tag{8}$$

The first-order condition (FOC) relative to *s* implies that the optimal load factor $l^* = dN/s^*F^*$ solves:

$$l^{*2}C_C'(l^*) = dC_K'\left(\frac{dN}{l^*}\right) \tag{9}$$

The choice of vehicle capacity therefore involves a trade-off between crowding costs and capital costs, which decrease and increase with *s*, respectively. If the capital cost function C_K is linear, from (9) the optimal load factor l^* is constant: vehicle size is chosen so that the product $s \cdot F$ (i.e. seat-kilometers) increases commensurately with the level of demand *N*. Otherwise the load factor may increase or decrease with *N* depending on the relative convexity (or concavity) of C_C and C_K.

[17] This result actually corresponds to a well-known result in the industrial organization literature, which is that if the cross partial derivative of inverse demand is null (as it is the case here, with $\partial^2 G/\partial N\partial F = \partial^2 G/\partial N\partial s = 0$), then a monopolist supplies quality using the same rule as if maximizing social welfare (Spence, 1975).

Regarding service frequency, two regimes arise depending on whether the operational constraint (4) is inactive or binding. In the normal regime, the constraint is inactive, meaning that the operator freely sets the service frequency. The combination of the two FOCs relative to *F* and *s* yields:

$$\frac{N}{F^2} C'_W\left(\frac{1}{F}\right) + \frac{N^2}{F^2}\delta'\left(\frac{N}{F}\right) C'_V\left(\delta\left(\frac{N}{F}\right)\right) = \left(\frac{1}{v} + \frac{1}{d_S}\delta\left(\frac{N}{F}\right) - \frac{N}{d_S F}\delta'\left(\frac{N}{F}\right)\right) C'_O\left(\frac{F}{v} + \frac{F}{d_S}\delta\left(\frac{N}{F}\right)\right) \qquad (10)$$

The choice of frequency involves this time a trade-off between waiting costs and in-vehicle costs versus operating costs. If the operating cost function C_O is linear, the optimal frequency increases less than linearly with demand. Otherwise, the relationship between *F* and *N* is once again undetermined as it depends on the relative convexity (or concavity) of C_W, C_V and C_O.

In the congested regime, the transit authority supplies the maximum feasible service frequency: $F^*(N) = \bar{F}(N)$. Accordingly, frequency strictly decreases with the level of demand *N*.

We illustrate the above results for the linear model (Figure 4, see Online Appendix A for details). We show that for this specification, both frequency and vehicle size increase less than linearly with demand at first in the normal regime. As demand exceeds a certain threshold ($N \geq \hat{N}$), the system enters the congested regime: the line is overcrowded, frequency gradually decreases while vehicle size increases supra-linearly with demand to compensate for the lower frequency (see Online Appendix A).

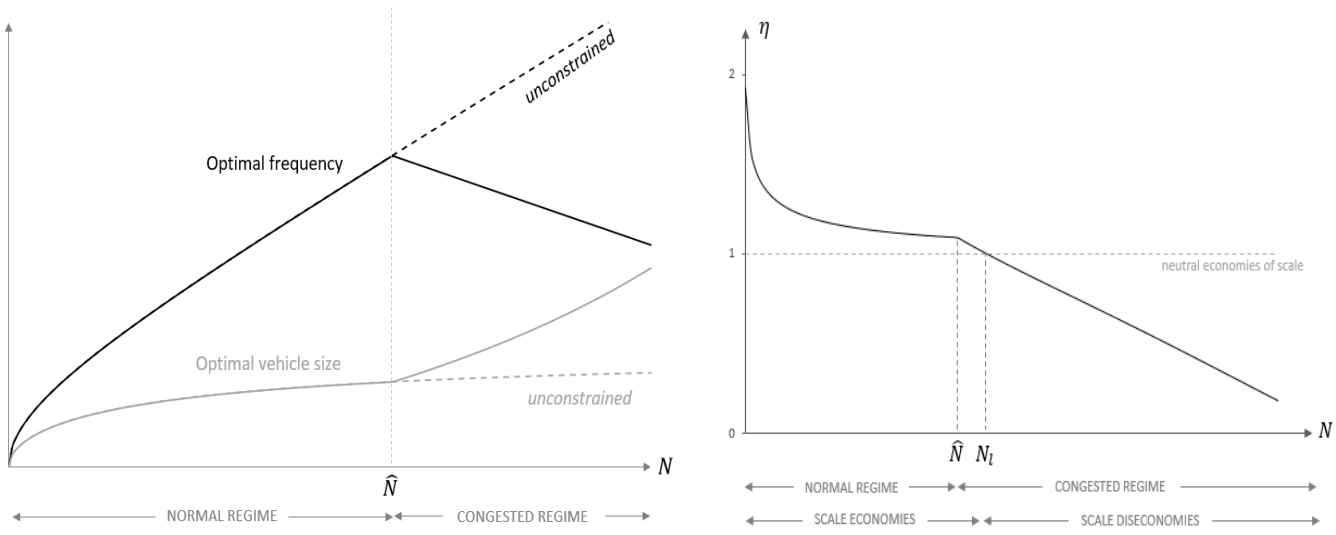


Figure 4: Optimal frequency and vehicle size (left) and degree of economies of scale (right) in the linear case

3.2.2. (Dis)economies of scale

We now turn to the crux of the paper, being the analysis of economies of scale – or economies of density to be specific as network size is fixed – and the influence of between-vehicle congestion.

Let $ASC \equiv SC/N$ denote the average social cost:

$$ASC(F,s,N) = C_W\left(\frac{1}{F}\right) + C_V\left(\delta\left(\frac{N}{F}\right)\right) + C_C\left(\frac{dN}{sF}\right) + \frac{C_K(sF)}{N} + \frac{C_0\left(\frac{F}{v} + \frac{F}{d_S}\delta\left(\frac{N}{F}\right)\right)}{N}. \quad (11)$$

We further consider three time horizons depending on which levers are available to the transit agency: in the short run, service frequency and vehicle size are both fixed; in the medium run the agency may adjust frequency, while finally in the long run both vehicle capacity and service frequency are flexible. We denote by $i \in \{s; m; l\}$ the horizon considered (short-run *s*, medium-run *m*, long-run *l*).

Proposition 1

For each time horizon, the provision of the public transit service is subject to economies of scale if $N < N_i^a$*, and to diseconomies of scale if* $N > N_i^b$*, with* $N_i^a < N_i^b$.

Whatever the time horizon, the system is characterized by economies of scale at low levels of demand, followed by diseconomies of scale at high levels of demand (the degree of economies of scale being undetermined in the general case for intermediate levels of demand). This similar pattern results from different economic forces at each time horizon, however.

In the short run (fixed service quality), economies of scale originate from the supply side, as the fixed production costs are split between more users. As demand increases, the negative user externalities - extra dwelling time, crowding - eventually prevail, however.

In the medium run (fixed vehicle size) and in the long run (adjustable frequency and vehicle size), economies of scale primarily find their source in user costs instead of production costs as previously. If the level of demand is low the frequency constraint (4) is inactive; the system is in the normal regime. The marginal user imposes the following externality to other agents:

$$N\frac{dASC^*}{dN} = -\frac{C_K}{N}\left(\frac{\eta_K - 1}{\eta_K}\right) - \frac{C_O}{N}\left(\frac{\eta_O - 1}{\eta_O}\right) - \frac{C_W}{\eta_W}. \quad (12)$$

where η_K, η_O and η_W represent the degree of economies of scale (with $\eta \equiv C/yC'$, $y \in \{K; O; W\}$) for capital costs, operating costs and waiting costs, respectively. As service quality improves with demand, the marginal user imposes on other users a positive externality– C_W/η_W (with $\eta_W \leq 1$ as C_W is convex). Although user costs do always entail economies of scale, overall (dis)economies of scale also depend on (dis)economies of scale on the supply side. Assume both capital costs and operating costs are subject to economies of scale ($\eta_K \geq 1$ and $\eta_O \geq 1$). From (12) we have $N \cdot dASC^*/dN < 0$. The normal regime is then always characterized by economies of scale, regardless of the convexity of the user cost functions. This includes as a specific case the linear specification $\eta_W = \eta_O = \eta_K = 1$, for which we find again the result of Mohring (1972) that the externality $N \cdot dASC^*/dN$ is equal to the average waiting cost $-C_W$. If $\eta_K \leq 1$ or $\eta_O \leq 1$, diseconomies of scale ($dASC^*/dN > 0$) may arise for intermediate values of *N*. For low values of *N*, the Mohring effect (the waiting externality – C_W/η_W) always prevails however, resulting in positive economies of scale.

As demand keeps increasing, the frequency constraint eventually becomes active and the railway line enters the congested regime. The marginal user now imposes the following externality:

$$N\frac{dASC^*}{dN} = -\frac{C_K}{N}\left(\frac{\eta_K - 1}{\eta_K}\right) - \frac{C_O}{N}\left(\frac{\eta_O - 1}{\eta_O}\right) - \frac{C_W}{\eta_W} + \lambda\left(\frac{\bar{F}}{N^2} - \frac{\bar{F}'}{N}\right), \quad (13)$$

where $\lambda > 0$ is the Lagrange multiplier associated to the frequency constraint, and $\bar{F}(N)$ the maximum feasible frequency (with $\bar{F}'(N) \leq 0$). The externality is the same as before with an additional term $\lambda(\bar{F}/N^2 - \bar{F}'/N) > 0$ that captures the deterioration of service frequency with each additional user. Accordingly, even if the system is always characterized by economies of scale in the normal regime, in the congested regime economies of scale may persist at first, but past a certain threshold, operations degrade to such an extent that diseconomies of scale will always occur.

Again, we illustrate these results for the linear model (Figure 4). In the normal regime ($N \leq \widehat{N}$), the system is characterized by economies of scale, as predicted. The degree of economies of scale decreases with demand, however: as demand increases, so does service frequency, waiting times fall leading the Mohring effect to become smaller and smaller. As the line becomes overcrowded ($N \geq \widehat{N}$), the degree of economies of scale falls even more sharply. It eventually gets lower than 1 past a certain threshold ($N \geq N_l$), implying diseconomies of scale.

3.2.3. Optimal pricing and subsidies

Consider first that the transit agency maximizes social welfare $SW^*(N) = SW(F^*(N), s^*(N), N)$. The first-order condition relative to the level of demand is $G(N^*) = MSC^*(N^*)$.This is the standard result that at optimum the marginal user benefit equals the marginal social cost. The optimal fare is:

$$\tau^* = MSC^*(N^*) - C_U(F^*(N^*), s^*(N^*), N^*). \quad (14)$$

From $MSC = ASC + N \cdot dASC/dN$ and $ASC = C_U + C_{TA}/N$, we can rewrite (14) as the standard first-best pricing rule (Small and Verhoef, 2007, eq (4.44)):

$$\tau^* = \frac{C_{TA}}{N^*} + N^*\frac{dASC^*}{dN}. \quad (15)$$

Corollary of Proposition 1

At each time horizon, the transit service is subsidized if $N^* < N_i^a$, and self-financed if $N^* \geq N_i^b$.

Let $\pi^* = C_{TA}/N^* - \tau^*$be the optimal subsidy per trip. Equation (15) implies $\pi^* = -N^* \cdot dASC^*/dN$. From Proposition 1 the optimal subsidy is positive at low levels of demand due to economies of scale, but negative if demand becomes too strong due to the Pigouvian principle and diseconomies of scale. In the latter case the service is self-financed, meaning that fares at least cover costs.

In the monopolistic case, the profit-maximizing fare corresponds to the previous pricing rule with the standard additional mark-up term $-N^e G'(N^e)$:

$$\tau^e = \frac{C_{TA}}{N^e} + N^e\frac{dASC^*}{dN} - N^e G'(N^e). \quad (16)$$

Proposition 2

Demand is always lower at the monopoly equilibrium than at optimum.

The monopoly and optimal solutions are characterized by the same provision rules for service quality, yet different demand levels. Monopolistic behavior involves raising the fare thus reducing demand relatively to the social optimum in order to maximize profit.

4. Extension: car competition

A frequent second-best rationale for subsidizing public transit is that car travel is underpriced in many cities around the world. We examine this argument in presence of public transportation congestion by considering that travelers can now choose between two modes: private car (C) and public transit (PT). Let N_C and N_{PT} denote the number of users, with $N = N_C + N_{PT}$ the total volume of travel demand. Road capacity is fixed, so that road congestion implies diseconomies of scale for car travel: we thus assume that the social cost function for the car mode $SC_C(N_c)$ is C^2 and convex. For ease of exposition, we further assume that for low volumes of demand the car mode is always the most efficient: $MSC_C(0) \leq \min_{F,s,N} MSC_{PT}(F, s, N)$.[18] The model setup is otherwise like in the general case (Section 3).

Consider the first-best social welfare maximization problem:[19]

$$\max_{s,F,N_{PT},N_C} \int_0^{N_C+N_{PT}} G(n)dn - SC_C(N_C) - SC_{PT}(F, s, N_{PT}). \qquad (17)$$
$$s.t. N_C \geq 0, N_{PT} \geq 0, F \leq \bar{F}(N_{PT})$$

From (17), it is straightforward to show that car competition does not change the optimal provision rules for public transit, only the optimal level of demand as stated by Proposition 4.

Proposition 4

At the first-best optimum, the demand for car and public transit are given by:

$$\begin{aligned} &G(N^*) = MSC_C(N^*), N_C^* = N^*,\ N_{PT}^* = 0 && \text{if } N^* \leq N_{PT>0} \\ &G(N^*) = MSC_C(N_C^*) = MSC_{PT}^*(N_{PT}^*) && \text{if } N^* \geq N_{PT>0} \end{aligned} \qquad (18)$$

As the total demand N^* increases, three successive patterns occur:

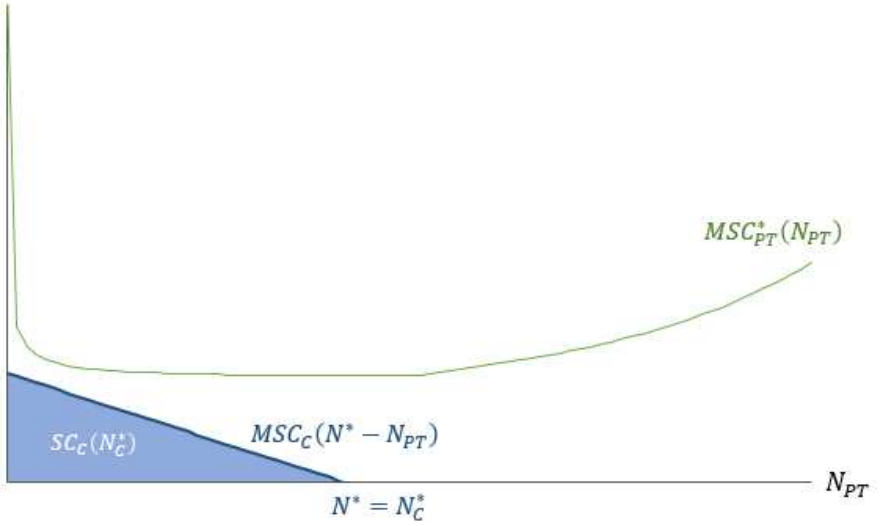
- *at first, all individuals use the car, and only N_C^* increases;*
- *then the number of public transit users increases, while the number of car users declines;*
- *eventually car users and public transit users both increase in number.*

First-best optimality implies equating marginal social costs across all modes used. If total demand is too low, the curve $MSC_C(N^* - N_{PT})$ does not intersect $MSC_{PT}^*(N_{PT})$. The cost of providing public transit is too high, and it is more efficient to use only the car (Figure 6). As total demand N^* increases, eventually $MSC_C(N^* - N_{PT})$ and $MSC_{PT}^*(N_{PT})$ intersect at two points, the second of which is a candidate for the optimal solution. It is only a candidate indeed, as a second condition for

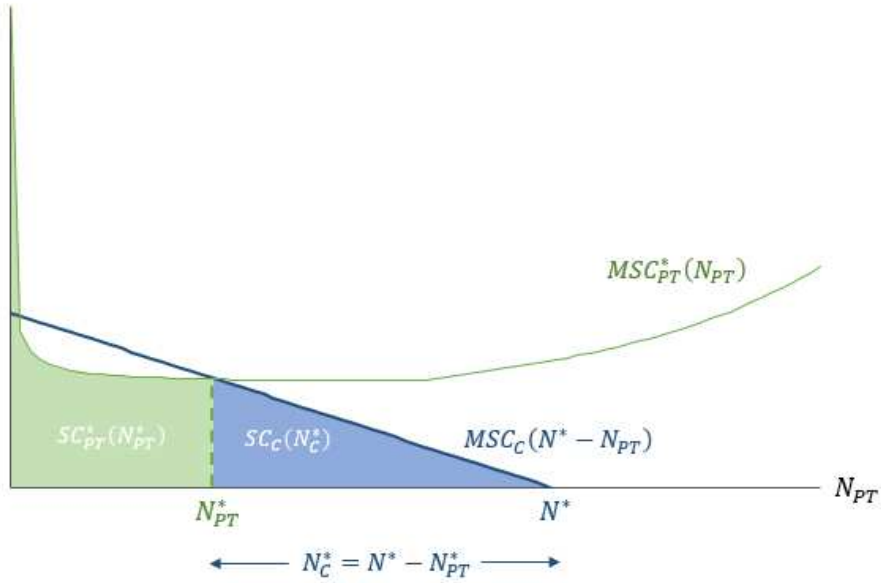
[18] This assumption has no influence on our results, and only allows us to rule out the case $N_c = 0$.

[19] Here we assume that the two modes are perfect substitutes and independent (no congestion across modes). These two assumptions are not central to the main findings reported in this subsection.

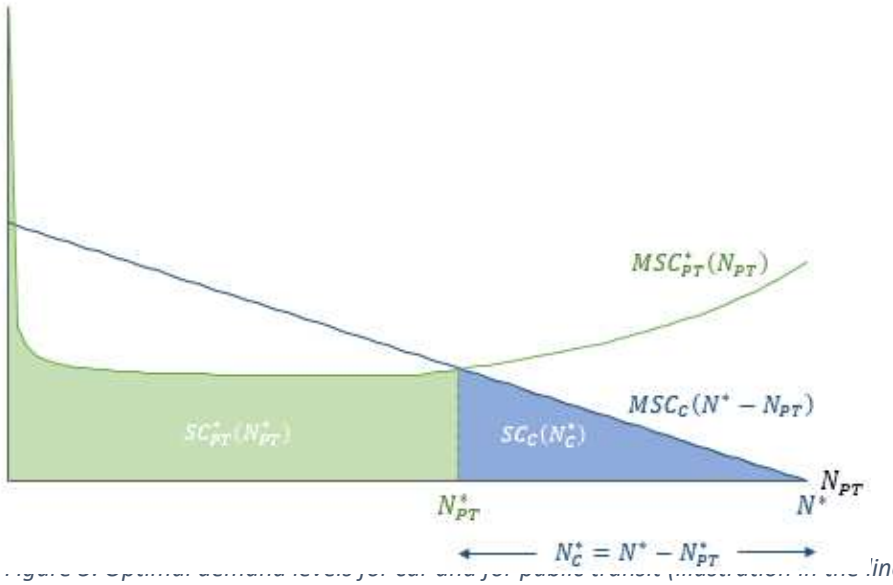
optimality is for the social cost $SC^*_{PT}(N^*_{PT})$ of providing public transit to N^*_{PT} users (green area in Figure 5) to be lower than the social cost of transporting the same users by car which, seeing that there are already $N^* - N^*_{PT}$ car users, is $SC_C(N^*) - SC_C(N^* - N^*_{PT})$ (quadrilateral under the blue curve in Figure 5). While this condition is not satisfied at first, eventually it becomes optimal to provide public transit. This causes the number of transit users to jump from 0 to $N^*_{PT} > 0$, while the number of car users drops from N^* to $N^* - N^*_{PT}$. From there, as total demand increases, the volume of public transit users N^*_{PT} increases. In the normal regime, this causes the marginal social cost $MSC^*_{PT}(N^*_{PT})$ to decrease. From $MSC_C(N^*_C) = MSC^*_{PT}(N^*_{PT})$, this implies that N^*_C must decrease as $MSC_C(.)$ is a strictly increasing function. In the congested regime, $MSC^*_{PT}(N^*_{PT})$ increases however, and thus so does N^*_C. As the public transport system becomes congested, the road mode becomes again a relevant alternative to transport additional users.



(a) Low total demand



(b) Intermediate total demand



(c) High total demand

Figure 3: Optimal demand levels for car and for public transit (illustration in the linear case)

Note: because $\forall N \geq 0\ MSC_C(0) < MSC^*_{PT}(N)$*, the curve* $MSC_C(N^* - N_{PT})$ *is always below* $MSC^*_{PT}(N_{PT})$ *at* $N_{PT} = N^*$.

Regarding pricing, the first-best solution is:

$$\tau_C^* = N_C^* \frac{dASC_C}{dN}(N_C^*), \tag{19}$$

$$\tau_{PT}^* = \frac{C_{TA}(F^*(N_{PT}^*), s^*(N_{PT}^*), N_{PT}^*)}{N_{PT}^*} + N_{PT}^* \frac{dASC_{PT}^*}{dN}(N_{PT}^*). \tag{20}$$

The optimal fare is the same as in the single-mode case, so that results regarding the self-financing of the system (Corollary of Proposition 1) still apply. Because it reduces public transportation demand,[20] car competition makes the transit service less likely to be congested, thus more likely to be subsidized. The optimal car tax is simply equal to the road externality $N_C^* \cdot dASC_C/\, dN$ (operating costs are not considered for this mode).

Assume now that car travel is not taxed ($\tau_C = 0$); car drivers only incur the private cost $C_U^C(N_C)$. The second-best solution, denoted by superscript $**$, is characterized by the following system:

$$\begin{aligned} &G(N^{**}) = C_U^C(N^{**}), N_C^{**} = N^{**},\ N_{PT}^{**} = 0 && \text{if } N^{**} \le N'_{PT>0} \\ &G(N^{**}) = C_U^C(N_C^{**}) = MSC_{PT}^*(N_{PT}^{**}) - \frac{-G'(N^{**})}{C'^C_U(N_C^{**}) - G'(N^{**})} \cdot \left(MSC_C(N_C^{**}) - C_U^C(N_C^{**})\right) && \text{if } N^{**} > N'_{PT>0} \end{aligned} \tag{21}$$

$$\tau_C^{**} = 0, \tag{22}$$

$$\tau_{PT}^{**} = \frac{C_{TA}}{N_{PT}^{**}} + N_{PT}^{**} \frac{dASC_{PT}^*}{dN}(N_{PT}^{**}) - \frac{-G'(N^{**})}{C'^C_U(N_C^{**}) - G'(N^{**})} \cdot \left(MSC_C(N_C^{**}) - C_U^C(N_C^{**})\right). \tag{23}$$

Compared to the first-best optimum, the transit fare is reduced by $-G'/\left(C'^C_U - G'\right)\left(MSC_C - C_U^C\right)$ to increase the competitiveness of public transit and compensate for car travel being underpriced. Because the ratio $-G'/\left(C'^C_U - G'\right)$ is strictly lower than one, the transit fare reduction is less than the implicit car subsidy $MSC_C - C_U^C$ resulting from the absence of road pricing. If demand is too strong, diseconomies of scale in public transit (captured by the term $N_{PT}^{**} \cdot dASC_{PT}^*/dN$) exceed this discount, however, resulting in a negative net subsidy: $\tau_{PT}^{**} > C_{TA}/N_{PT}^{**}$.

The above results are summarized in Figure 6. If total demand is low, the railway line is not economically sustainable; only the car mode is used. The latter being subject to diseconomies of scale, the average social cost steadily rises with demand. As demand further increases, public transit arises as a relevant alternative and the line is operated (with a subsidy). Economies of scale in public transit mitigate the increase of the average social cost. As congestion builds up, the public transit system eventually falls into diseconomies of scale. Subsidies are no longer necessary, but car use and the average social cost both steadily rise again as a result. In the second-best case, not taxing car travel leads to a greater modal share of the car despite a substantially larger subsidy of the railway service.

[20] From (18), if $N_{PT}^* > 0$ then $MSC_{PT}^*(N_{PT}^*) = G(N_{PT}^* + N_C^*)$. Considering that $G(N)$ strictly decreases with N and $N_C^* > 0$, this implies $MSC_{PT}^*(N_{PT}^*) < G(N_{PT}^*)$ and that the solution N_{PT}^* is lower than the solution N^* without the car mode (which satisfies $MSC_{PT}^*(N^*) = G(N^*)$).

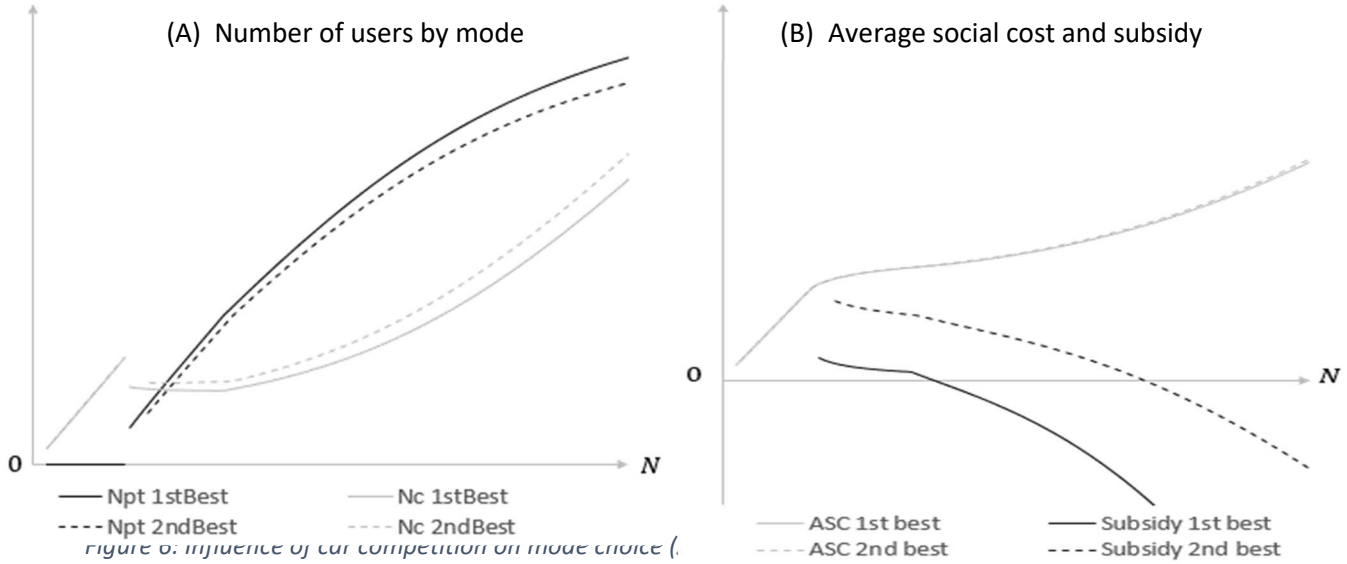


Figure 6: Influence of car competition on mode choice (

5. Empirical application

5.1. Data

In order to investigate empirically the existence of diseconomies of scale in public transit networks, we consider the Piccadilly Line, the second-longest tube line of the London Underground network. The Piccadilly line serves many of London's key tourist attractions as well as Heathrow airport (Figure A.2), making it the sixth busiest London tube line according to 2016/17 data from Transport for London (TfL). Due to an ageing rolling stock and insufficient frequency during peak times, the Piccadilly line faces recurrent overcrowding issues in its central section, which peak at King's Cross St Pancras – Russell Square interstation. A major capacity upgrade investment program is planned as part of the New Tube for London scheme in order to relieve congestion, making the Piccadilly line a prime candidate to illustrate the effects of overcrowding and possible ways to address them.

Table 2 reports key figures for the morning peak (7am - 10am) and hyperpeak (8am - 9am) periods, in the westbound direction.[21] The central section corresponds to the busiest part of the line, extending from Wood Green to Russell Square station. The Piccadilly line features an average service quality for a metro line, with a capacity of 684 users per vehicle[22] and a mean frequency of 22 trains/h during the morning peak. The frequency is slightly lower during the hyperpeak (21.7 trains/h), foreshadowing possible overcrowding issues. While the line is not very busy overall, it nears its maximum capacity in the central section, with an average load factor of 63% during the morning peak hour that rises to 82% during the hyperpeak.[23] This results from a substantially stronger demand in the central section - almost four times the average boarding rate (per km) of the

[21] In 2017, trip direction on the Piccadilly line during the morning peak period (7am – 10am) was split as follows: 56% westbound, 44% eastbound (Transport for London, 2017).
[22] The vehicle capacity is computed using a passenger standing density of 4pax/m².
[23] On the interstation Kings Cross – Russell square, the load factor even exceeds 100% during the hyperpeak.

whole line -, which is partly compensated for by the lower trip distances (6.07 km in the central section against 9.35 km for the whole line).

Table 2 : Key figures of the Piccadilly line (2017)

		Central section	Whole line
Length (km)		9.22	71.27
Number of interstations		8	51
Average interstation distance (km)		1.19	1.43
Vehicle capacity (users)		684	684
Frequency (trains/h)	*MP* [a]	22.0	22.0
	MH [b]	21.7	21.7
Mean travel distance (km)		6.07	9.35
Boarding (users/h.km)	*MP* [a]	1 560.1	425.1
	MH [b]	1 999.6	540.6
Average load factor	*MP* [a]	63%	27%
	MH [b]	82%	34%

[a] *MP: morning peak → 7am to 10am* [b] *MH: morning hyperpeak → 8am to 9am*

Source: Online Appendix A

Considering our objective to test for and investigate the effects of overcrowding, the application focuses on the central section of the Piccadilly line and on the morning hyperpeak period (8am – 9am). We consider linear specifications (see Online Appendix 1 for a detailed presentation of the linear case). The parameters used in the application of the linear model are grouped according to three categories (Table 3): technical, demand, and cost. The following paragraphs briefly discuss the parameter values and how they were computed. More extensive information is available in Online Appendix 2 .

The technical parameters describe the main characteristics of the line transportation technology. This includes interstation distance, free-flow commercial speed and vehicle capacity, which are readily available from TfL data. The minimum safe headway and the marginal dwelling time are estimated by regressing real supply (measured as the average of the five largest observed frequencies per every bracket of 25 tap-in per h.km) against the observed demand (measured by the per km validation rate). More specifically, we estimate the structural equation $F = F_0(1 - \delta d_M N)$ for the congested regime (Figure A. 3), using 2013 and 2014 frequency data and validation data collected on a hourly basis.[24] Because the value of d_M is known, this allows us to retrieve $H_0 = {F_0}^{-1}$ and δ. We find the minimum safe headway estimate $H_0 = 111.8\,s$, which corresponds to a maximum frequency of 32 trains/h. The marginal dwelling time is estimated to be 0.42 s per additional user.[25]

[24] To test for possible endogeneity issues, an alternate equation was estimated for the morning peak period only, for which the planned frequency is constant and equal to 24 trains/h for the whole period. This led to very similar results, strongly corroborating the absence of endogeneity issues (see discussion in online appendix).

[25] Lam et al. (1998) find a marginal dwelling time of δ = 0.037 s/user for the Hong-Kong mass rapid transit system, which converts here to δ = 0.082 s/user as metro carriages of the Piccadilly line consist of 18 double-doors (instead of 40 for Hong Kong). Puong (2000) finds in the case of the MBTA red line δ = 4.1 s/user/double-

Table 3: Parameter values

	Parameter	Value	*Source*
Technical			
d_S	Interstation distance (km)	1.19	*TfL – Interstation database*
s	Vehicle capacity (users)	684	*TfL - Rolling Stock Information Sheets*
v_F	Free-flow commercial speed (km/h)	40.89	*TfL – Interstation database*
H_0	Minimum safe headway (s)	111.8	*Authors' estimate from TfL validation and supply datasets*
δ	Marginal dwelling time (s/user)	0.43	*Authors' estimate from TfL validation and supply datasets*
Demand			
A	Maximum WTP (£/trip)	14.12	*Authors' estimate from RODS 2017*
B	Slope of WTP (£/trip/user.km^{-1}.h^{-1})	- 0.0040	*Authors' estimate from RODS 2017*
d	Mean trip length (km)	6.07	*Authors' estimate from RODS 2017*
Cost			
c_K	Capital cost parameter (£/seat.km)	0.0425	*TfL + Parry & Small (2009)*
c_O	Operating cost parameter (£/train.h)	1431.3	*TfL + Parry & Small (2009)*
α_W	Value of waiting time (£/h)	10.62	*Abrantes & Wardman (2011)*
α_V	Value of in-vehicle travel time (£/h)	7.33	*Abrantes & Wardman (2011)*
α_C	Maximum crowding penalty (£/trip)	1.65	*Whelan & Crockett (2009)*
μ	Marginal cost of public funds	0.3	*Kleven & Kreiner (2006)*

The linear demand function is estimated by crossing the observed demand level with the generalized price, assuming a generalized price elasticity of -0.75.[26] The mean trip length is estimated from the Rolling Origin and Destination Survey (RODS) 2017. The values of in-vehicle travel time and waiting time are borrowed from Abrantes and Wardman (2011), while the maximum crowding penalty was estimated by adapting the results of Whelan and Crockett (2009) to the case study.

Last, the operating cost and capital cost parameters are estimated using TfL financial reports, completed by cost parameters retrieved from Parry and Small (2009). The marginal cost of public funds is set to 0.3 (Kleven and Kreiner, 2006).

5.2. Baseline results

We first present the results for the hyperpeak period, in the medium run followed by the long run. Next, we contrast them with the results for the off-peak period. Although not presented here for the sake of concision, a sensitivity analysis confirms that our results are robust to the model specification and to the main parameter values (see Online Appendix A).

door, which again converts here to δ = 0.23 s/user. The greater marginal dwelling time estimate in this study is likely related to the high level of crowding. As a matter of fact, Puong (2000) empirically finds δ to significantly increase with the crowding level, as standees in the vehicle and/or on the platform hinder user transfer movements, causing each boarding and alighting to take more time.

[26] The generalized price elasticity of -0.75 is chosen as a central value from the empirical literature (Paulley et al., 2006). It is also very close to the value -0.8 reported by Parry and Small (2009) for peak rail travel in London.

5.2.1. Medium run

We first present the medium-run solution (fixed vehicle size). The discussion focuses on the optimal and monopolistic provision regimes, the MCPF solution being an intermediate between these two. Similarly, the tables below only report the main cases, the complete solutions being sent back to the Appendix (Table A.1 to A.3).

As expected, transit fares are greater under monopoly than at optimum (Table 4), while demand follows the opposite pattern. The ensuing high level of demand at optimum results in overcrowding, hence a lower frequency at optimum than under monopoly (22.7 against 25.4 trains/h, respectively). This contrasts with the standard result from the literature that frequency increases with demand and is thereby greater at optimum than under profit maximization.[27] All three components of the user cost are consequently greater at optimum: a lower frequency implies greater waiting costs and, combined to a stronger demand, longer boarding and alighting times as well as greater in-vehicle crowding. Conversely, stronger demand and lower frequency imply lower average operating costs at optimum. For all three provision regimes as well as the observed situation, the fare is set above the average operating cost, implying a negative subsidy regarding the first-best optimum. Incidentally, we find the observed transit fare (2.88£) to be close to the optimal one (2.58£), so that limited welfare gains are to be expected from pricing adjustments alone.

Table 4: Fare, economies of scale and welfare estimates

	Morning hyperpeak				Off-peak	
	Monopoly *medium run*	Optimum *medium run*	Optimum *long run*	*Observed*	Optimum *medium run*	*Observed*
Patronage (users/km.h)	1 312	2 077	2 423	*1 999*	866	*669*
Regime	*Normal*	*Congested*	*Congested*	*Congested*	*Normal*	*Normal*
Frequency (trains/h)	25.4	22.7	21.1	*21.7*	23.7	*20.5*
Vehicle capacity (users)	684	684	1767	*684*	684	*684*
User cost (£/trip)	2.33	3.15	2.61	*3.17*	1.57	*1.54*
waiting	*0.21*	*0.23*	*0.25*	*0.25*	*0.22*	*0.26*
in-vehicle travel time	*1.36*	*1.58*	*1.70*	*1.58*	*0.95*	*0.93*
crowding	*0.76*	*1.34*	*0.65*	*1.35*	*0.40*	*0.35*
Operating cost (£/trip)	1.41	0.87	1.13	*0.87*	0.90	*0.99*
vehicle capital costs	*0.56*	*0.32*	*0.65*	*0.55*	*0*	*0*
other operating costs	*0.85*	*0.55*	*0.48*	*0.32*	*0.90*	*0.99*
Price (£/trip)	6.50	2.58	1.74	*2.88*	0.68	*2.28*
Markup/tax (+) or subsidy (-)	5.08	1.71	0.61	*2.01*	-0.22	*1.29*
Waiting time (min.)	1.18	1.32	1.42	*1.38*	1.27	*1.46*
Travel time (min.)	11.16	12.92	13.94	*11.04*	7.76	*11.04*
Load Factor	46%	81%	40%	*82%*	24%	*21%*

[27] This is true for separable (in *N*, *F* and *s*) inverse demand functions. As discussed in Basso and Jara-Diaz (2010), a monopolist may oversupply frequency for more complex, non-separable inverse demand functions.

Economies of scale (£/trip)	0.21	-1.71	-0.61	*-1.49*	0.22	*0.28*
Social welfare (£/h)	10 140	12 258	13 313	*12 081*	2 567	*2 422*
Average social welfare (£/user)	7.73	5.90	5.49	*6.04*	2.96	*3.62*

Moving to the crux of the paper, we find substantial diseconomies of scale for both the optimum and MCPF cases and the observed situation. These are caused by the strong level of demand that leads to overcrowding and congested train operations. Moderate economies of scale persist under monopoly as the lower demand allows for normal train operations.

Breaking down the optimal subsidy shows that strong diseconomies of scale on the demand side are partly offset by economies of scale on the supply side (Table 5). Crowding is largely accountable for the negative subsidy, representing more than two thirds of the user externality. The Mohring effect (waiting externality) is on the other hand negligible due to the high frequency, contrasting with its preponderance in the theoretical literature. In order to assess the influence of the overcrowding effect, we relax the maximum frequency constraint and compute the new optimal subsidy. We find that failing to account for overcrowding entails substantial errors. Qualitatively, it leads to erroneous signs regarding the optimal subsidy and the overall user externality. Quantitatively, the absolute and relative magnitudes of the various externalities substantially differ depending on whether one considers the frequency constraint or not. With overcrowding, the crowding and travel time externalities become preponderant as frequency declines and may no longer be neglected. Regarding externalities on the supply side, while considering congestion between vehicles changes neither the sign nor the relative weight of each of the two elementary externalities, it does strongly affect their magnitude.

Table 5: Breakdown of the optimal subsidy

	Morning hyperpeak *medium run*		Morning hyperpeak *long run*		Off-peak *medium run*	
Between-vehicle congestion	*with*	*without*	*with*	*without*	*with*	*without*
Optimal subsidy (£/trip)	-1.71	0.14	-0.61	0.15	0.22	0.22
coming from						
waiting externality	-0.10	0.13	-0.13	0.13	0.19	0.19
travel time externality	-0.70	-0.02	-0.94	-0.05	-0.02	-0.02
crowding externality	-1.91	-0.04	0	0	-0.06	-0.06
capital cost externality	0.45	0.03	0	0	0	0
operating cost externality	0.54	0.04	0.47	0.07	0.11	0.11

5.2.2. Long run

Through adjustments in vehicle size, the transit agency is able to accommodate more demand in the long run. This results in lower fares, larger vehicles and greater demand levels than in the medium run (Table 7). The difference is especially salient at optimum, with an optimal vehicle size more than twice the current one. In the long run vehicle size is adjusted according to a target load factor that is both constant and the same for all provision regimes (Proposition 2), here 40%, hence the large

vehicle size at optimum in response to the stronger demand. This comes at the cost of a decrease in frequency: the optimal frequency falls to 21.1 trains/h in the long run, against 22.7 trains/h in the medium run. Frequency is again slightly greater under monopoly as the lower demand allows for normal operations, with 21.6 trains/h, against 25.4 trains/h in the medium run (Table A.2). By adjusting vehicle size, the transit agency is able to operate the line more efficiently. In the long run service provision is thus characterized by significantly lower diseconomies of scale at optimum, and slightly larger economies of scale under monopoly, all contributing to the lower transit fares (than in the medium run).

Despite the long run optimal subsidy being of the same sign as in the medium run, i.e. negative, its decomposition is substantially different (Table 8). As the long-run optimal provision rule states that vehicle capacity must be set to reach a (constant) target load factor, the crowding cost and capital cost per capita are equal and constant (Proposition 2), so that the corresponding externalities are zeroed. The travel time externality becomes the largest one (-0.94 £/user), again partly compensated by the operating cost externality, while the (negative) Mohring effect is slightly greater than in the short run. Again, accounting for congestion between vehicles leads to results that are quite different both qualitatively and quantitatively from the baseline model, though a lower gap in optimal subsidies relatively to the medium run.

5.2.3. Off-peak

The midday off-peak period (10a.m - 4p.m) allows contrasting the previous results with a lower demand case. Parameter values are kept the same as previously, except vehicle capital costs which are entirely assigned to the peak period and thus assumed to be 0, and demand parameters which are updated to match the off-peak level. Results are presented for the medium run only.

As demand is lower during the off-peak period, trains operate normally in all cases considered (Table 7, Table A.3). This leads to standard results from the literature, such as greater frequency and lower user costs at optimum than under monopoly, and to (mild) economies of scale for all provision regimes, implying a positive subsidy at optimum.

The analysis of the optimal subsidy falls likewise in line with the literature, with a dominating Mohring effect, followed by the operating cost externality (Table 8). Due to the lower demand levels, the crowding and travel time externalities are significantly lower than during the morning hyperpeak. Vehicle capital costs being entirely assigned to the morning peak period, there is no related externality. Here failing to account for overcrowding has obviously no effect as the line is not overcrowded during the off-peak period in the first place.

5.3. New Tube for London

Considering its strong usage and recurrent overcrowding issues during the morning peak period, the Piccadilly line is planned for an upgrade as part of a broader investment program called New Tube for London (NTfL). The investment objective regarding the Piccadilly line is to raise the total line capacity as well as to improve service quality through an increase in both vehicle size and frequency. The former will be achieved through the purchase of 94 new vehicles with enhanced carriage capacity.

The wider doors of the new vehicles will additionally allow to decrease boarding and alighting times. The NTfL program also includes upgrading the signaling system of the Piccadilly line in order to reduce the minimum safe headway, which combined to the improved boarding/alighting times will allow for higher frequencies during peak times. These investment decisions are in perfect line with our findings, that increasing vehicle capacity is welfare improving, but that with the current transport technology (in terms of minimum safe headway and boarding/alighting time) the line frequency would still be limited by the overcrowding, hence subject to diseconomies of scale.

Aiming to provide a first insight into the welfare effects of the NTfL program, we consider that it translates into the following changes for the Piccadilly line: 1) vehicle capacity s is expanded by 30%, 2) the unit boarding/alighting δ is decreased by 20%, and 3) the minimum safe headway is decreased to H_0 = 100 s (corresponding to a maximum frequency F_0 of 36 trains/h). We also consider a 20% demand increase at the corresponding time horizon (2025) for both the baseline and NTfL scenarios.

The results show that the increase in demand leads to a substantial degradation of service quality in the baseline scenario. Frequency decreases from 22.7 trains/h (Table 4) to 21.8 trains/h (Table 6), while the load factor rises from 81% to 92%. The line is subject to even greater diseconomies of scale as a result, as a marginal user imposes on other users an extra cost of -2.29£, against -1.71£ formerly. Compared to this do-nothing scenario, the NTfL program would as intended significantly improve both service frequency (from 21.8 to 25.1 trains/h) and comfort (the load factor falling from 92% to 72%). While it would still fall short from solving the overcrowding issue as it would attract yet more users, the NTfL program would limit diseconomies of scales (-32%) through greater operational efficiency, hence a substantial social welfare gain (+15%).

Table 6: Fare, economies of scale and welfare estimates of the NTfL program
(long run social optimum, morning hyperpeak)

	Baseline	**NTfL**
Patronage (users/km.h)	2 269	2 642
Regime	*Congested*	*Congested*
Frequency (trains/h)	21.8	25.1
Vehicle capacity (users)	684	889
User cost (£/trip)	3.41	2.93
waiting	*0.24*	*0.21*
in-vehicle travel time	*1.64*	*1.54*
crowding	*1.53*	*1.18*
Operating cost (£/trip)	0.79	0.75
vehicle capital costs	*0.28*	*0.28*
other operating costs	*0.51*	*0.47*
Price (£/trip)	3.08	2.30
Markup/tax (+) or subsidy (-)	2.29	1.56
Waiting time (min.)	1.38	1.19
Travel time (min.)	13.46	12.58

Load Factor	92%	72%
Economies of scale (£/trip)	-2.29	-1.56
Social welfare (£/h)	13 844	15 842
Average social welfare (£/user)	6.10	6.00

Conclusion

Our analysis suggests that very crowded lines face operational constraints regarding service frequency that lead to diseconomies of scale, as illustrated here for the Piccadilly subway line in London. When so, the fare should be set above the average operating cost, implying a negative subsidy (i.e., a tax). The key mechanism underpinning our findings is the presence of congestion between transit vehicles: beyond a certain level of demand, boarding and alighting takes so much time that frequency decreases because of trains sharing the same platform and of the minimum headway between successive trains. Adjusting vehicle capacity allows to accommodate more demand in the long run and thus to delay the occurrence of overcrowding, though only up to a certain extent.

Based on our model set-up, we show that out of the three frictions considered - between users, users and vehicles, and vehicles -, only between-vehicle congestion can lead to diseconomies of scale for user costs in the medium run and long run. While this result does not depend on the cost function specifications, it does depend on the key assumption that all three frictions scale either with the ratio between demand and frequency supply (N/F), or between demand and capacity supply (N/sF). These assumptions are common in theoretical and empirical works alike, so that they should not limit the extent of our findings unless new empirical evidence were to invalidate them. Similarly, the choice of a separable inverse demand function implies that the provision rules are the same for all three provision regimes (optimal, monopolistic, MCPF), so that ultimately differences in service quality are entirely driven by differences in the levels of demand. Opting for more complex, non-separable inverse demand functions could yield different results as established by Spence (1975). In light of the above, this work intends to stress that congestion between vehicles is a major source of diseconomies of scale for heavily used transit lines that should not be neglected and be addressed by appropriate policies (such as pricing or technological upgrades). This is shown theoretically using an analytical model that is otherwise always characterized by economies of scale, and empirically through the substantial corrections to the externalities estimates for the peak periods.

The analysis focuses on the case of a single line over a single time period (either peak or off-peak). Within a public transit network, the use of transit lines varies both in space (both between lines and along the lines, see Hörcher and Graham, 2018) and in time (between the peak and off-peak periods). Thus, our results suggest to enforce fare differentiation in order to shift demand away from the busiest lines toward less crowded time periods/transit lines. By doing so, diseconomies of scale on the congested lines would be partly if not fully compensated for by greater economies of scale on the less busy lines/time periods due to the increase in demand (Mohring effect). In the longer run, technological improvements to relax the frequency constraint, as in the New Tube for

London scheme, or network adjustments, such as designing alternate transit lines for the most popular OD pairs, could alleviate congestion, again mitigating diseconomies of scale (Jara-Díaz and Gschwender, 2003b). Because very busy lines often come with intensive land use along the line, land availability and land prices are often a significant hurdle to such infrastructure investments, however.

Among other caveats, the vehicle technology is deliberately represented in a simple fashion to keep the model tractable: a constant unit boarding/alighting time (implying that the number of openings remains constant and independent from vehicle capacity), yet no limit on vehicle capacity. The sensitivity analysis has shown that making boarding/alighting time a function of vehicle size delays the occurrence of overcrowding (as bigger vehicles handle boarding and alighting more efficiently), but does not change our main results, yet at the cost of much greater analytical complexity. Conversely, capping vehicle size (as train platforms cannot expand indefinitely) would strengthen diseconomies of scale by limiting the transit authority options to meet stronger demand - as shown by Hörcher (2017) - thus strengthening our main results. Finally, environmental externalities were not factored in the analysis. Again, these would lead to greater diseconomies of scale and social welfare losses if the transit line is congested, especially in presence of an unpriced road alternative.

To conclude, to reply to the question raised by Parry and Small (2009), "*Should urban transit subsidies be reduced?*", our model suggests that in some *not-so uncommon* cases, the answer should be "*yes*", and if heavily crowded urban transit systems remain subsidized, it should not be motivated by the usual rationales (economies of scale and underpriced car travel).

Acknowledgments

For useful comments, we would like to thank participants at the Annual Conference of the International Transportation Economics Association (ITEA) in Barcelona (Spain), June 2017, and in Paris (France), June 2019, at the S.T.E.F conference, Leuven (Belgium), June 2018, at the first Rencontres Francophones Transport Mobilité in Vaulx-en-Velin (France), June 2018, and at the 5th Winter Kraks Fond Workshop on Urban Economics in Copenhagen (Denmark), January 2020. We also thank Richard Arnott, Leonardo J. Basso, Daniel Hörcher, Sergio Jara-Díaz, Martin Koning, Alejandro Tirachini, and Jos Van Ommeren for insightful remarks, as well as the co-editor Owen Zidar and anonymous reviewers of the Journal of Public Economics.

This work was supported by the POLL-EXPO project, funded by the ADEME (grant number 1862C0012), and by the Lab Recherche Environnement funded by Vinci.

Declarations of interest: none.

References

Abrantes, P.A.L., Wardman, M.R., 2011. Meta-analysis of UK values of travel time: An update. Transportation Research Part A: Policy and Practice 45, 1–17. https://doi.org/10.1016/j.tra.2010.08.003

Badia, H., Estrada, M., Robusté, F., 2014. Competitive transit network design in cities with radial street patterns. Transportation Research Part B: Methodological 59, 161–181. https://doi.org/10.1016/j.trb.2013.11.006

Basso, L.J., Jara-Díaz, S.R., 2010. The Case for Subsidisation of Urban Public Transport and the Mohring Effect. Journal of Transport Economics and Policy 44, 365–372.

Basso, L.J., Silva, H.E., 2014. Efficiency and Substitutability of Transit Subsidies and Other Urban Transport Policies. American Economic Journal: Economic Policy 6, 1–33. https://doi.org/10.1257/pol.6.4.1

Benezech, V., Coulombel, N., 2013. The value of service reliability. Transportation Research Part B: Methodological 58, 1–15. https://doi.org/10.1016/j.trb.2013.09.009

Brueckner, J.K., Selod, H., 2006. The political economy of urban transport-system choice. Journal of Public Economics 90, 983–1005. https://doi.org/10.1016/j.jpubeco.2005.06.004

Canavan, S., Barron, A., Cohen, J., Graham, D.J., Anderson, R.J., 2019. Best Practices in Operating High Frequency Metro Services. Transportation Research Record 2673, 491–501. https://doi.org/10.1177/0361198119845356

Chang, S.K., Schonfeld, P.M., 1991. Multiple period optimization of bus transit systems. Transportation Research Part B: Methodological 25, 453–478. https://doi.org/10.1016/0191-2615(91)90038-K

Cominetti, R., Correa, J., 2001. Common-Lines and Passenger Assignment in Congested Transit Networks. Transportation Science 35, 250–267. https://doi.org/10.1287/trsc.35.3.250.10154

Daganzo, C.F., 2010. Structure of competitive transit networks. Transportation Research Part B: Methodological 44, 434–446. https://doi.org/10.1016/j.trb.2009.11.001

de Lapparent, M., Koning, M., 2016. Analyzing time sensitivity to discomfort in the Paris subway: an interval data model approach. Transportation 43, 913–933. https://doi.org/10.1007/s11116-015-9629-7

de Palma, A., Lindsey, R., Monchambert, G., 2017. The Economics of Crowding in Rail Transit. Journal of Urban Economics 101, 106–122. https://doi.org/10.1016/j.jue.2017.06.003

Fielbaum, A., Jara-Diaz, S., Gschwender, A., 2020. Beyond the Mohring effect: Scale economies induced by transit lines structures design. Economics of Transportation 22, 100163. https://doi.org/10.1016/j.ecotra.2020.100163

Fosgerau, M., 2009. The marginal social cost of headway for a scheduled service. Transportation Research Part B 43, 813–820. https://doi.org/10.1016/j.trb.2009.02.006

Gagnepain, P., Ivaldi, M., 2002. Incentive Regulatory policies: The Case of Public Transit Systems in France. RAND Journal of Economics 33, 605–629.

Hörcher, D., 2017. The economics of crowding in urban rail transport (Ph.D.). Imperial College London.

Hörcher, D., Graham, D.J., 2018. Demand imbalances and multi-period public transport supply. Transportation Research Part B: Methodological 108, 106–126. https://doi.org/10.1016/j.trb.2017.12.009

Jansson, K., 1993. Optimal public transport price and service frequency. Journal of Transport Economics and Policy 27, 33–50.

Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2020. Strategies for transit fleet design considering peak and off-peak periods using the single-line model. Transportation Research Part B: Methodological 142, 1–18. https://doi.org/10.1016/j.trb.2020.09.012

Jara-Díaz, S., Gschwender, A., 2003a. Towards a general microeconomic model for the operation of public transport. Transport Reviews 23, 453–469. https://doi.org/10.1080/0144164032000048922

Jara-Díaz, S., Gschwender, A., 2003b. From the Single Line Model to the Spatial Structure of Transit Services: Corridors or Direct? Journal of Transport Economics and Policy 37, 261–277.
Jara-Díaz, S., Tirachini, A., 2013. Urban Bus Transport: Open All Doors for Boarding. Journal of Transport Economics and Policy (JTEP) 47, 91–106.
Kleven, H.J., Kreiner, C.T., 2006. The marginal cost of public funds: Hours of work versus labor force participation. Journal of Public Economics 90, 1955–1973. https://doi.org/10.1016/j.jpubeco.2006.03.006
Kraus, M., 1991. Discomfort externalities and marginal cost transit fares. Journal of Urban Economics 29, 249–259. https://doi.org/10.1016/0094-1190(91)90018-3
Kraus, M., Yoshida, Y., 2002. The Commuter's Time-of-Use Decision and Optimal Pricing and Service in Urban Mass Transit. Journal of Urban Economics 51, 170–195. https://doi.org/10.1006/juec.2001.2242
Lam, W.H.K., Cheung, C.Y., Poon, Y.F., 1998. A study of train dwelling time at the Hong Kong mass transit railway system. Journal of Advanced Transportation 32, 285–295. https://doi.org/10.1002/atr.5670320303
Mohring, H., 1972. Optimization and Scale Economies in Urban Bus Transportation. The American Economic Review 62, 591–604. https://doi.org/10.2307/1806101
Oldfield, R.H., Bly, P.H., 1988. An analytic investigation of optimal bus size. Transportation Research Part B: Methodological 22, 319–337. https://doi.org/10.1016/0191-2615(88)90038-0
Parry, I.W.H., Small, K.A., 2009. Should Urban Transit Subsidies Be Reduced? The American Economic Review 99, 700–724. https://doi.org/10.2307/25592479
Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J., White, P., 2006. The demand for public transport: The effects of fares, quality of service, income and car ownership. Transport Policy 13, 295–306. https://doi.org/10.1016/j.tranpol.2005.12.004
Puong, A., 2000. Dwell time model and analysis for the MBTA red line (No. 02139–4307), Massachusetts Institute of Technology Research Memo.
Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. Routledge.
Spence, A.M., 1975. Monopoly, Quality, and Regulation. The Bell Journal of Economics 6, 417–429. https://doi.org/10.2307/3003237
Tirachini, A., Hensher, D.A., Jara-Díaz, S.R., 2010a. Restating modal investment priority with an improved model for public transport analysis. Transportation Research Part E: Logistics and Transportation Review 46, 1148–1168. https://doi.org/10.1016/j.tre.2010.01.008
Tirachini, A., Hensher, D.A., Jara-Díaz, S.R., 2010b. Comparing operator and users costs of light rail, heavy rail and bus rapid transit over a radial public transport network. Research in Transportation Economics, Reforming Public Transport throughout the World 29, 231–242. https://doi.org/10.1016/j.retrec.2010.07.029
Transport for London, 2017. TfL Rolling Origin and Destination Survey - London Datastore (See http://data.london.gov.uk/dataset/tfl-rolling-origin-and-destination-survey.).
Turvey, R., Mohring, H., 1975. Optimal Bus Fares. Journal of Transport Economics and Policy 9, 280–286.
Vickrey, W., 1980. Optimal transit subsidy policy. Transportation 9, 389–409. https://doi.org/10.1007/BF00177700
Whelan, G., Crockett, J., 2009. An investigation of the willingness to pay to reduce rail overcrowding, in: Proceedings of the First International Conference on Choice Modelling.
Yoshida, Y., 2008. Commuter arrivals and optimal service in mass transit: Does queuing behavior at transit stops matter? Regional Science and Urban Economics 38, 228–251. https://doi.org/10.1016/j.regsciurbeco.2008.01.004
Zhang, J., Yang, H., Lindsey, R., Li, X., 2019. Modeling and managing congested transit service with heterogeneous users under monopoly. Transportation Research Part B: Methodological. https://doi.org/10.1016/j.trb.2019.04.012

Appendix A – Notational glossary

Symbol	Variable
Technical	
d_S	Interstation distance (km)
F	Service frequency (trains/h)
H	Headway (h^{-1}/train)
$\bar{F}$	Maximum frequency (trains/h)
$\bar{H}$	Minimum headway (h^{-1}/train)
H_0	Minimum safe headway (h^{-1}/train)
s	Vehicle capacity (user/train)
t_W	Waiting time (h)
t_V	In-vehicle travel time (h)
t_D	Dwelling time (h)
l	Load factor (% of vehicle capacity)
v	Free-flow commercial speed (km/h)
H_0	Minimum safe headway (s)
δ	Marginal dwelling time (s/user)
Demand	
N	User arrival rate (user/h/km)
G	Generalized inverse demand function (£)
P	Inverse demand function (£)
A	Maximum WTP (£)
B	Slope of WTP (£/user.km^{-1}.h^{-1})
n_A	Number of users alighting per station and per vehicle
n_B	Number of users boarding per station and per vehicle
d	Mean trip length (km)
Cost	
X	Vehicle-kilometers supplied (train.km/km)
Z	Vehicle-hours supplied (train.h/km)
C_K	Capital cost function (£/seat.km)
C_O	Operating cost function (£/train.h)
C_{TA}	Total production cost function (£/train.h)
C_W	Waiting cost function (£/user)
C_V	In-vehicle time cost function (£/user)
C_C	Crowding cost function (£/user)
C_U	User generalized travel cost (£/user)
α_W	Value of waiting time (£/h)
α_V	Value of in-vehicle travel time (£/h)
α_C	Maximum crowding penalty (£/trip)
$\eta_.$	Degree of economies of scale (for cost indexed by .)
SW	Social welfare function (£/h)
SC	Social cost function (£/)
ASC	Average social cost function (£/user)
MSC	Marginal social cost function (£/user)
μ	Marginal cost of public funds
Pricing	
τ	Fare (£/user)
π	Subsidy (£/user)
Provision regime (superscript)	
*	Social optimum
^{e}	Monopoly equilibrium
$^\mu$	Social optimum with marginal cost of public funds

Appendix B – Additional figures

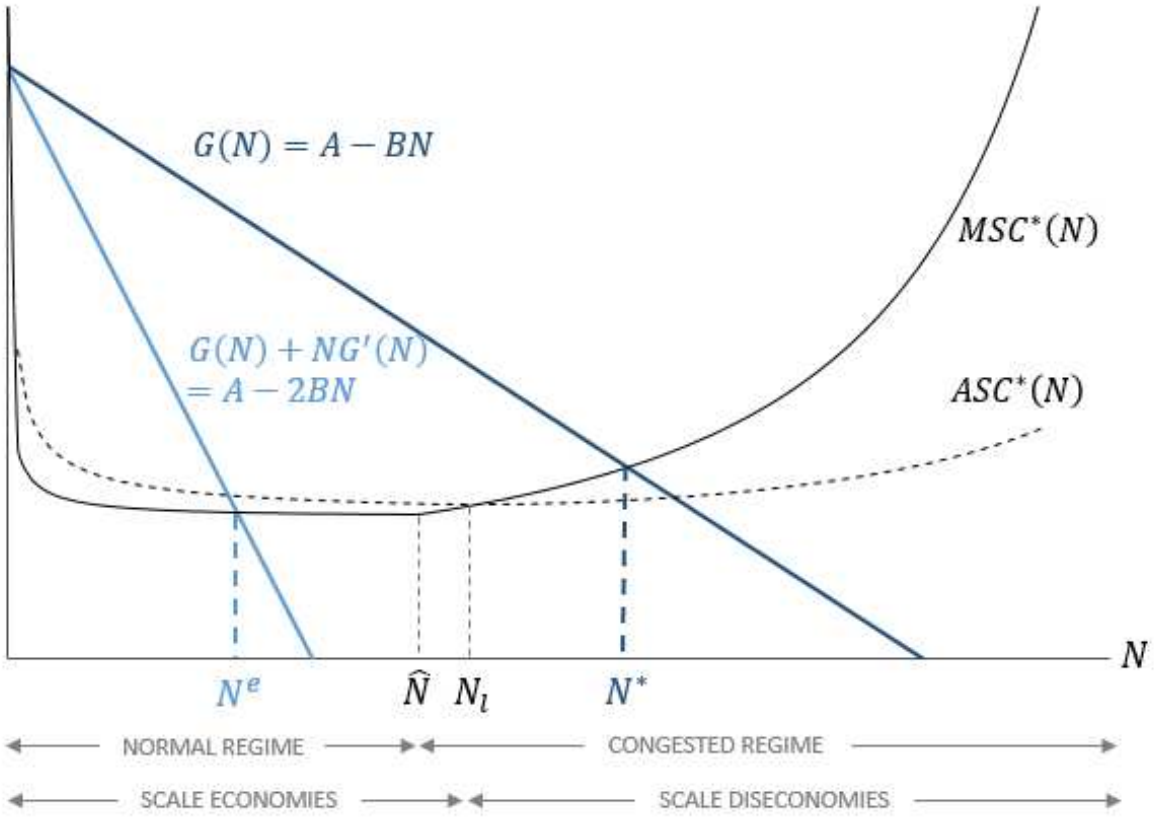


Figure A.1 : Optimal and monopolistic long-run levels of demand in the linear case

Note: here the optimal demand level falls within the hypercongested regime ($N^* > N_l$), while the monopolistic demand level falls within the normal regime ($N^e < \widehat{N}$). Changes in either A or B could lead to different situations, however.

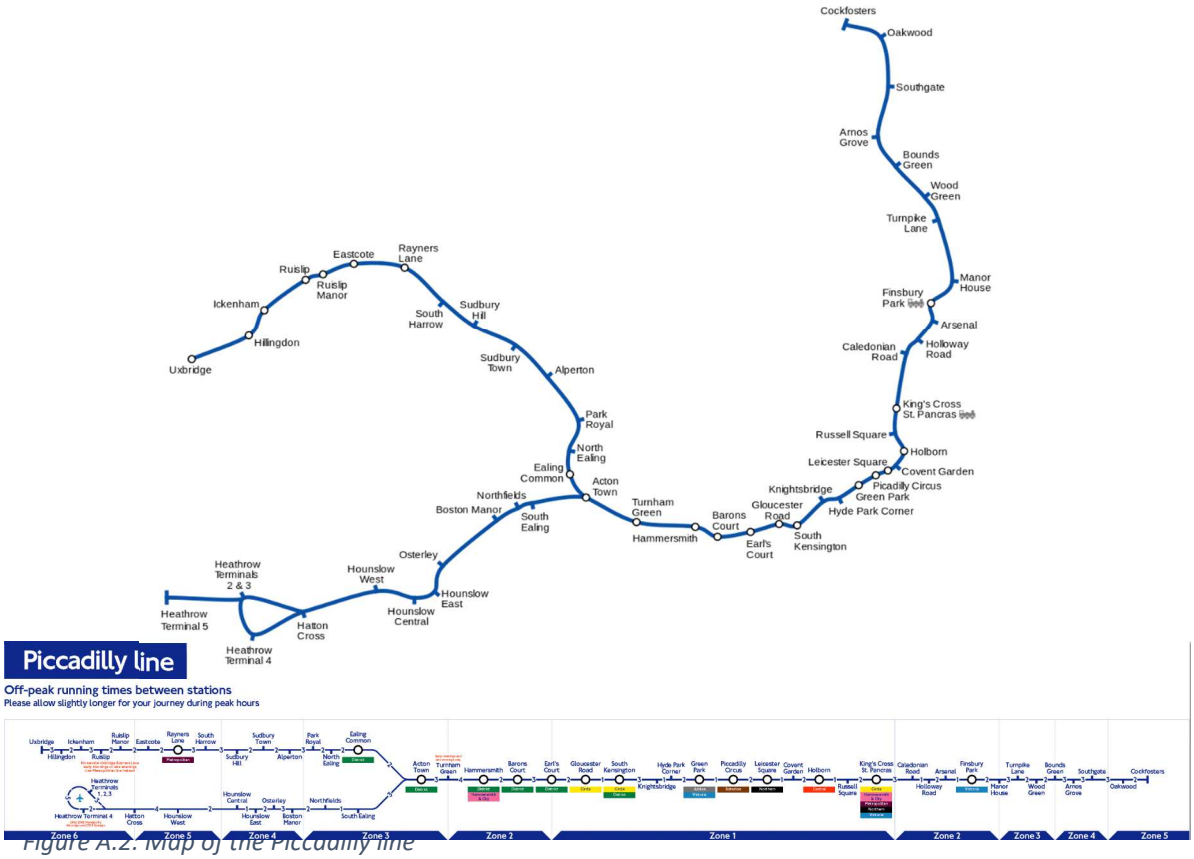


Figure A.2: Map of the Piccadilly line

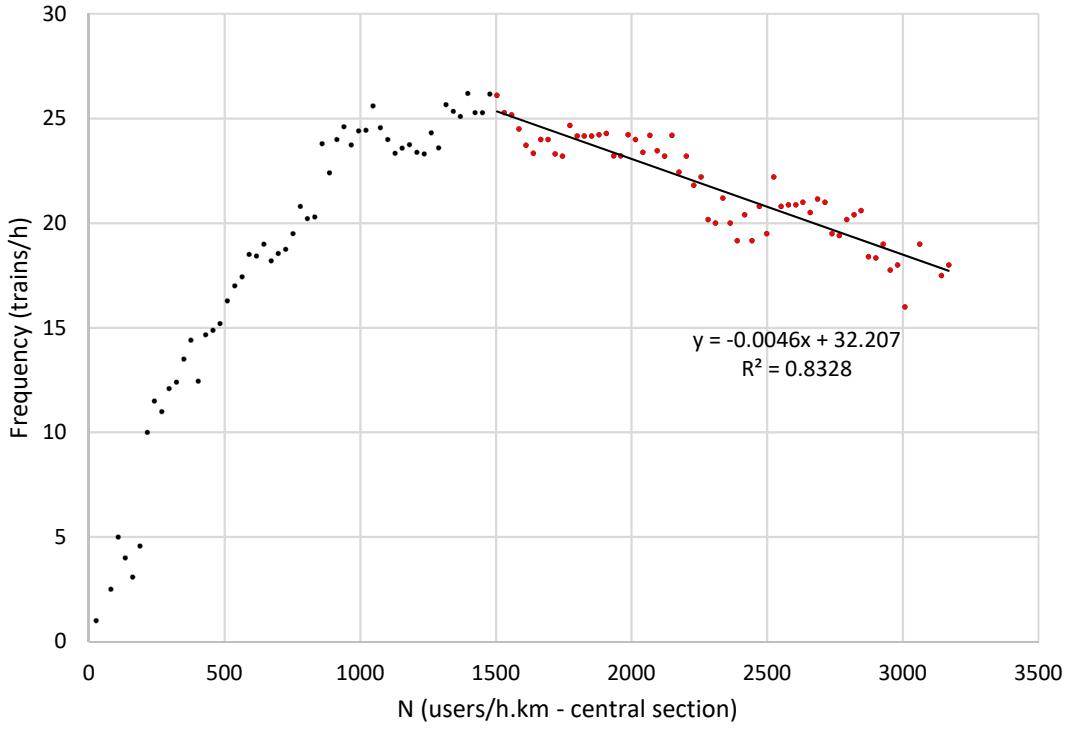


Figure A. 3: Relationship between real supply and observed demand levels

Note: one point corresponds to the average of the 5 highest frequencies for all observations in our dataset falling within a specific demand interval (every bracket of 25 tap-in per h.km). Observations of frequency and validations are at the daily level, hour per hour, for 2013 and 2014.

Table A.1: Fare, economies of scale and welfare estimates (medium run)

	Monopoly	**Optimum**	**MCPF**	**Observed**
Patronage (users/km.h)	1 312	2 077	1 848	1 999
Regime	*Normal*	*Congested*	*Congested*	*Congested*
Frequency (trains/h)	25.4	22.7	23.7	21.7
Vehicle capacity (users)	684	684	684	684
User cost (£/trip)	2.33	3.15	2.87	3.17
waiting	*0.21*	*0.23*	*0.22*	*0.25*
in-vehicle travel time	*1.36*	*1.58*	*1.50*	*1.58*
crowding	*0.76*	*1.34*	*1.14*	*1.35*
Operating cost (£/trip)	1.41	0.87	0.99	0.87
vehicle capital costs	*0.56*	*0.32*	*0.37*	*0.55*
other operating costs	*0.85*	*0.55*	*0.62*	*0.32*
Price (£/trip)	6.50	2.58	3.79	2.88
Markup/tax (+) or subsidy (-)	5.08	1.71	2.80	2.01
Waiting time (min.)	1.18	1.32	1.27	1.38
Travel time (min.)	11.16	12.92	12.31	11.04
Load Factor	46%	81%	69%	82%
Economies of scale (£/trip)	0.21	-1.71	-1.08	-1.49
Social welfare (£/h)	10 140	12 258	12 059	12 081
Average social welfare (£/user)	7.73	5.90	6.53	6.04

Table A.2: Fare, economies of scale and welfare estimates (long run)

	Monopoly	**Optimum**	**MCPF**	**Observed**
Patronage (users/km.h)	1 320	2 423	2 058	1 999
Regime	*Normal*	*Congested*	*Congested*	*Congested*
Frequency (trains/h)	21.6	21.1	22.7	21.7
Vehicle capacity (users)	939	1767	1389	684
User cost (£/trip)	2.31	2.61	2.46	3.17
waiting	*0.25*	*0.25*	*0.23*	*0.25*
in-vehicle travel time	*1.41*	*1.70*	*1.57*	*1.58*
crowding	*0.65*	*0.65*	*0.65*	*1.35*
Operating cost (£/trip)	1.40	1.13	1.21	0.87
vehicle capital costs	*0.65*	*0.65*	*0.65*	*0.55*
other operating costs	*0.74*	*0.48*	*0.56*	*0.32*
Price (£/trip)	6.48	1.74	3.36	2.88
Markup/tax (+) or subsidy (-)	5.08	0.61	2.15	2.01
Waiting time (min.)	1.39	1.42	1.32	1.38
Travel time (min.)	11.58	13.94	12.86	11.04
Load Factor	40%	40%	40%	82%
Economies of scale (£/trip)	0.25	-0.61	-0.23	-1.49
Social welfare (£/h)	10 225	13 313	12 960	12 081
Average social welfare (£/user)	7.74	5.49	6.30	6.04

Table A.3: Fare, economies of scale and welfare estimates (off-peak, medium run)

	Monopoly	**Optimum**	**MCPF**	**Observed**
Patronage (users/km.h)	430	866	703	669
Regime	*Normal*	*Normal*	*Normal*	*Normal*
Frequency (trains/h)	13.4	23.7	19.8	20.5
Vehicle capacity (users)	684	684	684	684
User cost (£/trip)	1.67	1.57	1.59	1.54
waiting	*0.40*	*0.22*	*0.27*	*0.26*
in-vehicle travel time	*0.93*	*0.95*	*0.94*	*0.93*
crowding	*0.35*	*0.40*	*0.38*	*0.35*
Operating cost (£/trip)	1.01	0.90	0.93	0.99
vehicle capital costs	*0*	*0*	*0*	*0*
other operating costs	*1.01*	*0.90*	*0.93*	*0.99*
Price (£/trip)	3.78	0.68	1.85	2.28
Markup/tax (+) or subsidy (-)	2.77	-0.22	0.93	1.29
Waiting time (min.)	2.24	1.27	1.51	1.46
Travel time (min.)	7.61	7.76	7.72	11.04
Load Factor	21%	24%	23%	21%
Economies of scale (£/trip)	0.40	0.22	0.27	0.28
Social welfare (£/h)	1 874	2 567	2 469	2 422
Average social welfare (£/user)	4.35	2.96	3.51	3.62

Appendix C - Proofs

Section 3. Model

Frequency constraint

The headway constraint writes $H \geq H_0 + \delta_0 + \delta(NH)$, with $H_0 + \delta_0 > 0$. Both sides strictly increase with H. If the two curves never intersect $\forall H \in \mathbb{R}^+$, there is no feasible headway for the level of demand *N*. Otherwise, let $\bar{H}(N)$ denote the first (strictly) positive solution to $H = H_0 + \delta_0 + \delta(NH)$. The headway constraint implies $H \geq \bar{H}(N)$. Then let $\bar{F}(N) \equiv \bar{H}(N)^{-1}$. We can rewrite the headway constraint in terms of frequency:

$$F \leq \bar{F}(N). \tag{24}$$

If $N = 0$, the frequency constraint is $F \leq F_0$ where $F_0 = (H_0 + \delta_0)^{-1}$. As *N* increases, the right-hand side of the headway constraint strictly increases with *N*, whereas the left-hand side remains constant. This implies that $\bar{H}(N)$ strictly increases with *N*, and thus that $\bar{F}(N)$ strictly decreases with *N*.

As *N* increases, two cases arise depending on whether the equation $H = H_0 + \delta_0 + \delta(NH)$ always has a solution or not.

If yes, because the function δ is unbounded, then $\lim_{N \to +\infty} \bar{H}(N) = +\infty$ and $\lim_{N \to +\infty} \bar{F}(N) = 0$.

If not, there exists N_{max} so that the equation $H = H_0 + \delta_0 + \delta(NH)$ has a solution if $N < N_{max}$, and none if $N > N_{max}$. Because $\bar{F}(N)$ strictly decreases with *N*, it can then either converge toward 0 or toward a strictly positive value. Convergence toward 0 occurs when the curve H is the asymptote of the curve $H_0 + \delta_0 + \delta(N_{max}H)$ for $N \to N_{max}$.The case of a strictly positive value occurs when the curves H and $H_0 + \delta_0 + \delta(N_{max}H)$ intersect at a single point $\bar{H}(N_{max})$, meaning that the curves are tangent at $H = \bar{H}(N_{max})$. In the latter case, the implicit function theorem yields: $(N\delta' - 1)\bar{F}' = \bar{F}\delta'$. Using the fact that the two curves are tangent at $\bar{H}(N_{max})$, then we have $N_{max}\delta' = 1$. As $N \to N_{max}$, the right-hand side is strictly positive, while on the left-hand side $(N\delta' - 1)$ tends toward 0. This implies that $\lim_{N \to N_{max}} \bar{F}'(N) = -\infty$.

Provision of service quality in the monopoly and MCPF regimes

Let us first consider that the transit agency maximizes profit: $\Pi(F, s, N) = N.P(N) - C_{TA}(F, s, N)$. From $SC(F, s, N) = N \cdot C_U(F, s, N) + C_{TA}(F, s, N)$, we can rewrite the profit function as:

$$\Pi(F, s, N) = N \cdot G(N) - SC(F, s, N). \tag{25}$$

Profit maximization is here formally equivalent to social welfare maximization, except that the aggregate gross user benefit $\int_0^N G(n)dn$ is replaced by the generalized gross revenue $N \cdot G(N)$. Provision rules for frequency and vehicle size are therefore the same at the monopoly equilibrium and at optimum: $s^*(N) = s^e(N)$ and $F^*(N) = F^e(N)$. Service quality is yet not necessarily the same, inasmuch as the monopolistic and optimal levels of demand N^e and N^* may differ.

The same result holds true for the MCPF regime, as the objective function of the transit authority $SW^{\mu}(F,s,N) = SW(F,s,N) + \mu\Pi(F,s,N)$ is actually a linear combination of the previous two.

Optimal service quality

The Lagrangian corresponding to the social cost minimization problem under the frequency constraint writes: $\mathcal{L} = SC(F,s,N) + \lambda\big(F - \bar{F}(N)\big)$. The first-order condition (FOC) relative to vehicle size *s* is:

$$\frac{\partial \mathcal{L}}{\partial s} = \frac{\partial SC}{\partial s} = -\frac{dN}{s^2 F} C_C'\left(\frac{dN}{sF}\right) + \frac{F}{N} C_K'(sF) = 0,$$

$$\Leftrightarrow \qquad l^{*2} C_C'(l^*) = d C_K'\left(\frac{dN}{l^*}\right). \tag{26}$$

where for reminder $l^* = dN^*/s^*F^*$ is the optimal load factor.

The FOC relative to service frequency is: $\partial\mathcal{L}/\partial F = \partial SC/\partial F + \lambda = 0$. This gives:

$$\frac{1}{F^2} C_W'\left(\frac{1}{F}\right) + \frac{N}{F^2}\delta' \cdot C_V'\left(\delta\left(\frac{N}{F}\right)\right) + \frac{dN}{sF^2} C_C'\left(\frac{dN}{sF}\right) = \frac{s}{N} C_K'(sF) + \frac{1}{N}\left(\widehat{t_V} - \frac{N}{F}\widehat{t_V}'\right) C_O'(F \cdot \widehat{t_V}) + \lambda,$$

where $\widehat{t_V}(N/F) = 1/v + \delta(N/F)/d_S$ is the travel time per km (i.e. the inverse of the commercial speed).

Assume first that the frequency constraint is not binding: $\lambda = 0$. By using the FOC with respect to vehicle size, the above equation simplifies to:

$$\frac{N}{F^{*2}} C_W'\left(\frac{1}{F^*}\right) + \frac{N^2}{F^{*2}}\delta'\left(\frac{N}{F^*}\right) C_V'\left(\delta\left(\frac{N}{F^*}\right)\right) = \left(\widehat{t_V}\left(\frac{N}{F^*}\right) - \frac{N}{F^*}\widehat{t_V}'\left(\frac{N}{F^*}\right)\right) C_O'\left(F^* \cdot \widehat{t_V}\left(\frac{N}{F^*}\right)\right) \tag{27}$$

If the frequency constraint is binding then $F^* = \bar{F}(N)$.

Proposition 1

Consider first the short-run optimum (*F* and *s* constant). The derivative of the average social cost is:

$$\frac{dASC}{dN} = \frac{1}{F}\delta' \cdot C_V'(\delta) + \frac{d}{sF} C_C' - \frac{C_K}{N^2} + \frac{\delta'}{d_S}\frac{C_O'}{N} - \frac{C_O}{N^2} \tag{28}$$

Each of the first three terms strictly increases with *N*, and $-C_K/N^2$ converges toward -∞ if *N* tends toward 0. Thus, the sum of the first three terms strictly increases with *N*, and is negative for low values of *N* and positive for high values of *N*. For the last two terms, there are two possibilities. If their sum is positive, this implies that $dASC/dN$ remains positive for high values of *N* (as the first three terms are already positive). For low values of N, the terms in $-1/N^2$ prevail, meaning that $dASC/dN$ is indeed negative. Conversely if the sum of the last two terms is negative, C_0/N decreases with *N*. Using the fact that C_0/N always remains positive, we also show that $dASC/dN$ is strictly negative if *N* is low, and strictly positive if *N* is high.

In the short run, the system is characterized by economies of scale ($dASC/dN < 0$) if demand is lower than N_s^a, and diseconomies of scale ($dASC/dN > 0$) if it is greater than N_s^b, with $N_s^a < N_s^b$.

Consider next the medium-run optimum. In the normal regime the constraint is inactive. We can use the envelope theorem to derive the medium-run optimal average social cost, which yields:

$$\frac{dASC^*}{dN} = \frac{1}{F}\delta' C'_V(\delta) + \frac{d}{sF}C'_C - \frac{C_K}{N^2} - \frac{C_O}{N^2} + \frac{\delta'}{d_S}\frac{C'_O}{N}.$$

Let $\eta = AC/MC = (C/y)/C'$ denote the degree of scale economies. Using the FOC relative to frequency, the above equation can be rewritten as:

$$N\frac{dASC^*}{dN}(N) = -\frac{C_K}{N}\left(\frac{\eta_K - 1}{\eta_K}\right) - \frac{C_O}{N}\left(\frac{\eta_O - 1}{\eta_O}\right) - \frac{C_W}{\eta_W}.$$

Considering that C_W is convex, we have $\eta_W \leq 1$ and $-C_W/\eta_W \leq -C_W < 0$. In the normal regime, the system is always characterized by economies of scale on the demand side. Overall economies of scale do depend on the degree of economies of scale on the supply side (through the first two right terms), however. If demand is very low ($N \to 0^+$), the average social cost tends toward +∞ as either frequency tends toward 0, in which case C_W tends toward +∞, or frequency remains above a strictly positive value, in which case C_O/N tends toward +∞. Considering the continuity and regularity of the average social cost, this implies that it necessarily strictly decreases in the neighborhood of 0.

In the congested regime, the constraint is active. Using the envelope theorem for constrained optimization yields:

$$N\frac{dASC^*}{dN}(N) = -\frac{C_K}{N}\left(\frac{\eta_K - 1}{\eta_K}\right) - \frac{C_O}{N}\left(\frac{\eta_O - 1}{\eta_O}\right) - \frac{C_W}{\eta_W} + \lambda\left(\frac{\bar{F}}{N^2} - \frac{\bar{F}'}{N}\right) \quad (29)$$

This is the same as before, with the addition of the last term that represents the effect of the headway constraint. Considering that $\lambda > 0$ and $\bar{F}' > 0$, this last term is always strictly positive and contribute to diseconomies of scale.

As *N* increases, two cases arise depending on whether *N* can tend toward +∞ or only toward N_{max} (see proof of frequency constraint). If *N* can tend toward +∞, the frequency decreases toward 0, which causes the average user cost to tend toward +∞ (due to the waiting cost). Because the production cost is positive, this means that the average social cost also tends toward +∞, and that it necessarily increases beyond a certain range. If *N* can only tend toward N_{max}, the argument is the same if the frequency decreases toward 0. If the frequency tends toward a strictly positive value $\bar{F}(N_{max}) > 0$, then all the terms in (29) are bounded except for the last term $-\lambda\bar{F}'/N$, which converges toward +∞ (as $\lim_{N \to N_{max}} \bar{F}'(N) = -\infty$). This implies that $\lim_{N \to N_{max}} N \cdot dASC^*/dN = +\infty$.

Again, the system is characterized in the medium run by economies of scale if demand is lower than N_m^a, and diseconomies of scale if it is greater than N_m^b, with $N_m^a < N_m^b$.

The long-run case is analogous to the medium run case, in particular equation (29) still holds.

Proposition 2

The profit maximization problem is: $\max_N N \cdot G(N) - SC^*(N)$, The first-order condition becomes: $G(N) + N \cdot G'(N) = MSC^*(N)$. Compared to the optimal pricing rule $G(N) = MSC^*(N)$, the left-

hand side includes an additional negative term. Because $G(N) + N \cdot G'(N)$ strictly decreases with N and is always strictly below the curve $G(N)$, it follows that $N^e \leq N^*$.

Proposition 4

As in Section 3, we first solve for the optimal frequency and vehicle size as functions of N_{PT}. By doing so we can rewrite the maximization problem (17) as:

$$\max_{N_{PT},N_C} \int_0^{N_C+N_{PT}} GC(n)dn - SC_C(N_C) - SC^*_{PT}(N_{PT}). \quad (30)$$

$$s.t. \begin{cases} N_C \geq 0 \\ N_{PT} \geq 0 \end{cases}$$

There are three possible cases: 1) $N_C^* = 0$, 2) $N_{PT}^* = 0$, and 3) $N_C^* > 0$ and $N_{PT}^* > 0$.

In case 1, there are no car users, $N_{PT}^* = N^*$and the marginal social cost is $MSC^*_{PT}(N^*) > 0$. Considering that $MSC_C(0) < MSC^*_{PT}(N^*)$, keeping the total number of users N^* constant, the social welfare can be increased by switching an infinitesimal quantity $\varepsilon > 0$ of users from public transport to the road. This means that case 1 is absurd, and that we always have $N_C^* > 0$.

In case 2, the public transport mode is not used ($N_{PT}^* = 0$), hence $N_C^* = N^*$. The maximization problem involves only one variable, with the first-order condition: $GC(N_C^*) = MSC_C(N_C^*)$.

In case 3, the two modes are used: $N_C^* > 0$ and $N_{PT}^* > 0$. By combining the two first-order conditions we get: $GC(N^*) = MSC_C(N_C^*) = MSC^*_{PT}(N_{PT}^*)$. This implies that N_{PT}^* satisfies:

$$MSC_C(N^* - N_{PT}^*) = MSC^*_{PT}(N_{PT}^*). \quad (31)$$

The LHS strictly decreases with N_{PT}^*, and strictly increases with N^*. We also have $\forall N \in \mathbb{R}^+$, $MSC_C(0) < MSC^*_{PT}(N)$. For low values of demand ($N^* < N_0$), the curve of the marginal social cost of car remains strictly below that of public transit and the two never intersect. We are always in case 2: it is more efficient to transport all users by car. If $N^* \geq N_0$, there exists at least one solution to $MSC_C(N^* - N_{PT}^*) = MSC^*_{PT}(N_{PT}^*)$. The question is then whether the interior solution using the two modes $SC_C(N^* - N_{PT}^*) + SC^*_{PT}(N_{PT}^*)$ does perform better than the car-only corner solution $SC_C(N^*)$. For $N^* = N_0$ the curves $MSC_C(N^* - N_{PT}^*)$ and $SC^*_{PT}(N_{PT}^*)$ are tangent, hence $SC_C(N^*) < SC_C(N^* - N_{PT}^*) + SC^*_{PT}(N_{PT}^*)$ (see Figure 6 for a graphical intuition). Because SC_C is convex, as N^* increases the cost of the car solution $SC_C(N^*)$ increases faster than $SC_C(N^* - N_{PT}^*) + SC^*_{PT}(N_{PT}^*)$, so that at some point $N^* = N_{PT>0}$ the interior solution becomes always the most efficient.

We now show the second part of Proposition 4. Consider an increase in N^*. If $N^* < N_{PT>0}$, we are in case 2, $N_{PT}^* = 0$ and $N_C^* = N^*$, meaning that N_C^* increases with N^*.

If $N^* > N_{PT>0}$, then $N_{PT}^* > 0$. In Equation (31), an increase in N^* causes the LHS to increase so that the marginal social cost of the car exceeds that of public transit. This causes N_{PT}^* to increase with N^* in order to equilibrate the two marginal social costs. If the public transit system is characterized by economies of scale, then the marginal social cost $MSC^*_{PT}(N_{PT}^*)$ decreases, meaning that $MSC_C(N_C^*) = MSC^*_{PT}(N_{PT}^*)$ also decreases. Because MSC_C is a strictly decreasing function, it follows

that N_C^* decreases with N^*. If $N_{PT}^* \geq N_l^b$ however, i.e. the system is characterized by diseconomies of scale, then $MSC_{PT}^*(N_{PT}^*)$ increases this time as N^* and N_{PT}^* increase, implying that N_C^* increases again.