Congestion, diseconomies of scale and subsidies in urban public transportation

Nicolas Coulombel, Guillaume Monchambert

Congestion, diseconomies of scale and subsidies in urban public transportation*

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

November 21, 2019

Working paper

Abstract

Subsidization of urban public transportation systems is often motivated by economies of scale and/or second-best considerations (underpriced road alternative). We model a public transportation system subject to frictions between users, users and vehicles, and vehicles. We derive the monopolistic and optimal provisions of supply. We show that if demand exceeds a first threshold, the system enters a congested regime and service frequency decreases. If demand exceeds a second threshold, the public transit system operates under diseconomies of scale, calling for a Pigovian tax instead of a subsidy. This finding, which goes against Mohring's classical rule (1972), holds with an untolled road alternative. We estimate the model for the London Piccadilly lane and find evidence of substantial diseconomies of scale during the morning peak, questioning current subsidy policies for the busiest transit lines.

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

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

*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 and at the first Rencontres Francophones Transport Mobilité in Vauls-en-Velin (France), June 2018. We also thank Leonardo J. Basso, Daniel Hörcher, Martin Koning and Alejandro Tirachini for insightful remarks.

[a] LVMT, UMR-T 9403, Ecole des Ponts, IFSTTAR, UPEM, Champs-sur-Marne, France (nicolas.coulombel@enpc.fr).

[b] LAET, University of Lyon, Université Lyon 2, Lyon F69007, France (g.monchambert@univ-lyon2.fr).

1. Introduction

Urban public transportation systems are heavily subsidized in many cities across the world (Table 1). The economic literature advances two main rationales for doing so (Parry and Small, 2009). First, public transit systems operate under economies of scale. While economies of scale may arise from production costs (Farsi et al., 2007; Ripplinger and Bitzan, 2018; Viton, 1992), a primary source of economies of scales in public transit is related to user costs, as shown by Mohring (1972): if public transit supply increases with patronage, a rise in demand will cause the average waiting time of users to diminish through an increase in service frequency, a phenomenon commonly referred to as the *Mohring effect*. Second, because car travel is typically underpriced relatively to the external costs that it generates - partly due to the unpopularity of road pricing (see De Borger and Proost, 2012) - public transportation is subsidized in order to support mode switching and limit car use, as a second best solution (Adler and van Ommeren, 2016; Anderson, 2014; Glaister and Lewis, 1978; Nelson et al., 2007). Vickrey (1980) mentions a third rationale, which is addressing the special needs for transit by the underprivileged, such as people with disabilities or low-income individuals who are unable or cannot afford to drive or access other forms of transportation.

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

Country	City	Public Transport 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	Sidney	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).

While the relevance of each rationale may vary depending on the city characteristics (road transport and public transport supply, road pricing, socio-economic conditions...), empirical studies typically find substantial scale economies associated to the Mohring effect (Nash et al., 2001; Nelson

et al., 2007; Parry and Small, 2009; Savage, 2010), justifying its central role in the theoretical literature.[1] Yet, a key assumption underlying the Mohring effect is that service frequency increases with demand. As a matter of fact, the optimal frequency increases with the square root of demand – the so-called *square root principle* - in the simplest version of Mohring's model, or following a modified square root formula when introducing additional features such as variable boarding/alighting time or crowding (Jara-Díaz and Gschwender, 2003a). However, in current urban context, public transit patronage has increased to such an extent that this assumption does not hold anymore: there are increasing cases of very congested lines for which the headway between two trains (or subways, buses, etc.) increases if demand is too strong, as a result of too many users seeking to board or to alight at each station. Moreover, most studies cited above largely ignore user crowding costs, which are yet a crucial consumption externality characterizing urban public transportation (de Palma et al., 2017).

This paper investigates the effect of congestion on economies of scale in public transportation, including implications in terms of pricing and subsidies. We develop an analytically tractable model that captures several key features of urban public transit congestion (in-vehicle crowding, effects on dwelling time and frequency). We then study the effect of increasing levels of demand on the provision of service quality (frequency, vehicle size/capacity) and on economies of scale for two provision regimes: monopolistic (profit-maximizing) and optimal (social welfare maximizing). Finally, we calibrate and apply the model to the London Piccadilly lane, and provide insights regarding the welfare effects of the New Tube for London (NTfL) scheme.

Our model builds on the theoretical microeconomic framework developed by Mohring (1972) and later extended by Jansson (1980, 1993), among others.[2] This framework has been widely applied to investigate public transit operations, economies of scale, optimal pricing rules and associated subsidies (Basso and Silva, 2014; Jansson, 1980, 1993; Mohring, 1972). We extend this framework by explicitly including three types of frictions: between users, between users and vehicles, and between vehicles. Following Kraus (1991), we model frictions between users as crowding. The crowding cost increases linearly with the in-vehicle occupancy rate, as typical in the literature (see e.g. de Palma et al., 2017). Frictions between users and vehicles are represented by considering that the boarding and alighting time (i.e. the dwelling time) increases linearly with the number of users, as in Mohring (1972). 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 transportation operations are characterized by economies of scale only up to a certain level of demand. If demand is too strong, the severity of crowding causes the marginal social cost of an extra passenger to exceed the average social cost, implying diseconomies of scale. Scale diseconomies arise in both the short-run (frequency and vehicle size are kept fixed), medium-run (adjustable frequency only) and long-run (adjustable frequency and vehicle size). Inasmuch as they imply lower equilibrium demand levels than at optimum, scale diseconomies still occur but are reduced

[1] While the theoretical validity of the Mohring effect was questioned at some point by van Reeven (2008), the controversy was short lived as two answers to his paper showed his results to be mostly spurious, definitely confirming if need be the Mohring effect (Basso and Jara-Díaz, 2010; Savage and Small, 2010).

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

under monopolistic provision or with a marginal cost of public funds, however. The direct corollary of these findings is that the optimal subsidy is negative when the system becomes too congested due to very strong demand (in relation to the transportation technology), which is a classical Pigovian result. Our findings are shown to be robust in presence of an unpriced substitute transportation mode, meaning that second-best pricing does not necessarily imply subsidies to public transportation users. The application to the Piccadilly line in London confirms empirically that not accounting for congestion leads to substantially overestimating the Mohring effect during the peak period, thus misestimating the sign of scale economies (from negative to positive). During the off-peak period the preponderance of the Mohring effect is reasserted, as the lower crowding level leads to normal operations and the usual economies of scale.

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 be reduced yet to subside when accounting for congestion. While all three frictions contribute to diseconomies of scale, we show that in the long run (under adjustable frequency and vehicle size) diseconomies of scale only arise in the presence of congestion between vehicles, underlining the importance of accounting for this specific mechanism.[3] Our model is also analytically tractable, allowing for clear-cut results as opposed to other works. Finally, we are - to the best of our knowledge - also the first to provide empirical evidence of diseconomies of scale regarding user costs and the social cost in urban public transportation.

The analysis focuses on the case of non-planning users.[4] The phenomena addressed in this paper (severe crowding and between-vehicle congestion) mostly concern very busy transit lines with short headways. Users are therefore much more likely not to plan under such conditions (Fosgerau, 2009; Jansson, 1993). First considered by Oldfield and Bly (1988), the effect of congestion on waiting times (as in denied boarding because the vehicle is full) is not considered here. This would involve moving from a steady-state to a dynamic model capturing queues on platforms, as in Kraus and Yoshida (2002) or Yoshida (2008) who use the bottleneck model, yet at the cost of much greater analytical complexity. Considering this effect would further increase diseconomies of scale as users would have to bear the additional cost of waiting for the next train as demand becomes too strong and frequency deteriorates. Conversely spatial effects linked to network design and line density (Jara-Díaz and Gschwender, 2003b) are expected to curb congestion and thus diseconomies of scale in the longer run, provided that the transit authority is able to meet additional demand with new transit infrastructures (which in many cities proves increasingly difficult due to the rising population and soar in land prices).

Our results call for a clear review of subsidies schemes in congested urban transportation systems. Considering that (dis)economies of scale are strongly related to the demand level, our findings also provide additional support for fare differentiation and peak pricing.

[3] We thereby generalize previous findings from Hörcher (2017) and Tirachini et al. (2010), who also find scale diseconomies yet in more restrictive contexts (hard constraints on frequency and vehicle size 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 schedule (Fosgerau, 2009).

2. A model of transit line with congestion

Consider a transit line, with stations evenly spaced and separated by an interstation distance d_M. Without loss of generality, we will refer to it as a railway line in the remainder of the paper.

Following Mohring (1972), we study the steady state of a one kilometer long route segment over a given time period - typically one hour during the morning peak period. Users are "non-planning", i.e. they do not look at the schedule, so that in each station new users arrive at a constant rate over time. The user arrival rate per hour and per kilometer of railway line is denoted by *N*, which measures the level of demand. To simplify matters, trip length is assumed constant and equal to *d*.

2.1. Transportation technology

Service frequency is noted *F* (trains/hour), while the headway is noted $H \equiv F^{-1}$. The service is assumed to be regular (constant headways) and reliable (the service always adheres to the schedule).[5] Vehicle size (capacity) is noted s and is supposed to be the same for all vehicles.

From the model assumptions – constant trip distances, arrival rates, and headways – the number of users alighting (n_A) and boarding (n_B) is the same at each station and for each train. It is given by: $n_A = n_B = d_M\ N/F$. As passengers stay onboard for d/d_M stations, the vehicle load is equal to dN/F. The level of crowding l_C is measured by the load factor, defined as the vehicle load over capacity:

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

The total travel time is the sum of access time t_A, waiting time t_W and in-vehicle travel time t_V. Interstation distance (and line density) remaining constant throughout the analysis, access time can be assumed to be 0 without loss of generality. For regular headways, the average waiting time is half the headway:

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

As in Mohring (1972), the commercial speed v is given by: $d_M/v = d_M/v_V + \delta_0 + \delta_A\ n_A + \delta_B\ n_B$, where d_M is the interstation distance, v_V the cruising speed between stops, δ_0 a fixed additional time per stop, and δ_A and δ_B are the unit alighting and boarding time per passenger, respectively.

Assumption 1

The unit alighting and boarding times δ_A and δ_B are independent of vehicle size s.

Assumption 1 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 (by optimizing the seat configuration, making smaller seats…), or by expanding the size of each car (horizontally or vertically) while leaving the number of openings constant (such as in switching from single-decker to double-decker trains).

[5] See Benezech and Coulombel (2013) for the case of unreliable transit services with non-planning users.

Let $\delta = \delta_A + \delta_B$, and $v_F = (1/v_V + \delta_0/d_M)^{-1}$ be the free-flow speed, i.e. the commercial speed without users in the system. This leads to: $1/v = 1/v_F + \delta N/F$. Finally, in-vehicle travel time is:

$$t_V = d\left(\frac{1}{v_F} + \delta\frac{N}{F}\right). \tag{3}$$

In Mohring's original model, service frequency is not upper bounded: as demand keeps increasing, optimal frequency tends toward infinity. Yet, train circulation is subject to operational constraints. First, the headway cannot physically be lower than the dwelling time. Moreover, regulators enforce an additional minimum safe headway H_0 between trains to limit collisions. This implies the following condition on the headway: $H \geq H_0 + \delta_A\, n_A + \delta_B\, n_B$. Through substitutions, this rewrites as:

$$F \leq F_0(1 - \delta d_M N), \tag{4}$$

where $F_0 = {H_0}^{-1}$ is the free-flow maximum frequency (without users in the system). The technological constraint (4) sets the maximum feasible frequency. As demand increases, more time is required for allowing passengers to alight and to board, fewer trains can pass and the maximum frequency declines.

2.2. Demand

Demand is characterized by a linear inverse demand function $G(N)$, where $G(N)$ denotes the reservation generalized price of the N^{th} user:[6]

$$G(N) = A - BN. \tag{5}$$

The standard reservation price is given by $P(N) = G(N) - C_U$, *i.e.* by subtracting from the reservation generalized price $G(N)$ the user travel cost C_U, which is here specified as follows:

$$C_U(t_W, t_V, l_C) = \alpha_W t_W + \alpha_V t_V + \alpha_C l_C. \tag{6}$$

α_W and α_V are respectively the values of waiting time and in-vehicle travel time, both expressed in monetary terms, and α_C is the crowding penalty factor. For model tractability, the crowding penalty is assumed to be independent of in-vehicle travel time t_V.[7]

Using (1), (2) and (3), the user travel cost can be rewritten as a function of frequency, vehicle size and demand:

$$C_U(F, s, N) = \frac{\alpha_W}{2F} + \alpha_V d\left(\frac{1}{v_F} + \delta\frac{N}{F}\right) + \alpha_C d\frac{N}{sF}. \tag{7}$$

There are two sources of externality regarding user costs: an additional user increases in-vehicle travel time (by increasing dwelling time), as well as in-vehicle crowding.

[6] Linear demand functions can be supported by considering homogeneous users with a quadratic utility function (as in Silva and Verhoef, 2013), or heterogeneous users with a linear utility function but uniformly distributed reservation utility levels (as in Basso and Jara-Díaz, 2010).

[7] The assumption of a constant crowding penalty as opposed to one linear in in-vehicle travel time has been empirically supported by the study of De Lapparent and Koning (2016), among others.

2.3. Production costs

Transit operations imply production costs which are assumed to be ultimately supported by the transit agency.[8] As the model represents a single line and does not account for variable line density, we overlook infrastructure costs and focus on operating costs instead (as in Parry and Small, 2009). Operating costs include vehicle capital costs and other operating costs, which depend on two primary inputs, vehicle-kilometers (noted X) and vehicle-hours (noted Z), as well as on vehicle size s. We consider the following specification:

$$C_{TA}(X, Z, s) = c_K sX + c_O Z. \tag{8}$$

The cost $C_{TA}(X, Z, s)$ per kilometer of route is the sum of capital costs $c_K sX$ and of operating costs $c_O Z$. Capital cost capture the depreciation of vehicles, which is assumed proportional to the distance travelled and to the vehicle size. Operating costs are based on vehicle hours, and correspond to the cost of drivers and other time-based operating costs.[9]

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

$$C_{TA}(F, s, N) = c_K sF + \frac{c_O}{v_F} F + c_O \delta N. \tag{9}$$

3. Optimal service quality and pricing

Two provision regimes are considered in this section: optimal and monopolistic. Consider first the monopolistic case where the transit authority maximizes profit: $\Pi(F, s, N) = NP(N) - C_{TA}(F, s, N)$. Let $SC(F, s, N) = N\, C_U(F, s, N) + C_{TA}(F, s, N)$ denote the social cost of the system (per kilometer of steady state route and per hour). The profit function rewrites:

$$\Pi(F, s, N) = N\, G(N) - SC(F, s, N). \tag{10}$$

Because the first term $N\, G(N)$ - which corresponds to a gross generalized revenue - is independent of *F* and *s*, the profit maximization corresponds to a bi-level optimization problem: 1) for a given *N*, choosing *F* and *s* so as to minimize the social cost, and 2) optimal choice of *N* at the upper level.

Next, consider now that the transit authority supplies service quality (frequency, vehicle capacity) and sets the fare in order to maximize social welfare:

$$SW(F, s, N) = \int_0^N G(n)dn - SC(F, s, N). \tag{11}$$

[8] 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.

[9] 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".

The social welfare maximization problem is similar to the profit maximization problem, except that the gross generalized revenue $N\,G(N)$ is replaced by the aggregate gross user benefit $\int_0^N G(n)dn$. Consequently, both problems involve choosing *F* and *s* so as to minimize the social cost for a given *N*, and the provision rules for frequency and vehicle size are the same at equilibrium and at optimum: $s^*(N) = s^e(N)$ and $F^*(N) = F^e(N)$.[10] Service quality is yet not necessarily the same across regimes, inasmuch as the equilibrium and optimal levels of demand N^e and N^* may differ.

Accordingly, we first discuss the optimal provision of service quality (*F* and *s*) at the lower level, then the monopolistic and optimal pricing rules (associated to N^e and N^*) at the upper level .

3.1. Optimal service quality

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

$$\min_{F,s} SC(F,s,N) \quad s.t.\ F \le F_0(1-\delta d_M N) \tag{12}$$

Regarding train frequency, two regimes arise depending on whether the frequency constraint (4) is inactive (normal regime) or binding (congested regime).

Proposition 1

In the normal regime, the transit authority supplies service frequency so as to equate production costs with the sum of waiting, dwelling and crowding costs.

In the congested regime, the transit authority supplies the maximal feasible frequency.

The choice of frequency involves the usual trade-off between production costs and variable user costs (excluding in-vehicle costs, here assumed constant). If demand is too strong, however, boardings and alightings take so much time that it induces congestion between vehicles (similarly to bus bunching). Frequency declines as a result, a phenomenon that we will refer to as “overcrowding”.

Regarding the choice of capacity, larger vehicles reduce crowding, but involve higher capital costs. The outcome of this trade-off is provided by Proposition 2.

Proposition 2

The transit authority supplies vehicle size so as to equate crowding costs with capital costs.

From Proposition 2, the optimal load factor is constant and equal to:

$$l_C^* = \sqrt{\frac{c_K d}{\alpha_C}}. \tag{13}$$

[10] 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).

Vehicle occupancy increases with the capital cost parameter c_K and decreases with the crowding cost parameter α_C, as expected.

Combining Propositions 1 and 2 yields the optimal provision of service quality (Table 1, see Proof in Appendix).

Table 1: *Optimal provision of service frequency and vehicle size*

	Normal regime ($N \le \widehat{N}$)	Congested regime ($\widehat{N} \le N \le N_{max}$)
$F^*(N)$	$\sqrt{\frac{v_F}{c_O}\left(\frac{\alpha_W}{2}N + \alpha_V \delta d N^2\right)}$	$F_0(1-\delta d_M N)$
$s^*(N)$	$\sqrt{\frac{c_O}{v_F}\frac{\alpha_C dN/c_K}{\frac{\alpha_W}{2}+\alpha_V \delta d N}}$	$\sqrt{\frac{\alpha_C d}{c_K}}\frac{N}{F_0(1-\delta d_M N)}$

The threshold N_{max} denotes the maximum level of demand for which a steady state solution exists, with $N_{max} = 1/\delta d_M$ (see Appendix). The threshold demand level $\widehat{N}$ marks the separation between the normal regime and the congested regime. It is the first positive solution to:

$$\frac{\alpha_W}{2}\widehat{N} + \alpha_V \delta d \widehat{N}^2 = \frac{c_O}{v_F}{F_0}^2\left(1-\delta d_M \widehat{N}\right)^2. \quad (14)$$

The RHS of (14) zeroes in $N = N_{max}$, implying $\widehat{N} < N_{max}$ (see Figure A.1 in Appendix). From (14), one can further show that (see also Figure A.1):

- an increase in either of the two technological parameters δ (the unit boarding/alighting time) and $H_0 = {F_0}^{-1}$ (the minimum safe headway) increases the risk of overcrowding (lower $\widehat{N}$);
- increasing interstation distance d_M also raises the risk of overcrowding;
- other parameters influence positively (vehicle speed v_F, demand parameters α_W, α_V and d) or negatively (operating cost parameter c_O) the risk of overcrowding inasmuch as they push the transit authority to raise frequency.

We now study the behavior of F^* and s^* with respect to N. Let η_F and η_S be the demand elasticity of service frequency and vehicle size, respectively.

Proposition 3

In the normal regime, an increase in demand leads to an increase in both frequency and vehicle size, with $0 \le \eta_S \le 1/2 \le \eta_F \le 1$.

In the congested regime, an increase in demand leads to an increase in vehicle size but to a decrease in frequency, with $\eta_S > 1$ *and* $\eta_F < 0$.

In the normal regime ($N \le \widehat{N}$), the optimal frequency and vehicle capacity both increase with demand, as expected (Figure 1). Frequency follows a modified square root formula (Table 1), which is actually

exactly the same as in Jansson (1980).[11] For low levels of demand, frequency is low and the waiting time effect prevails. As demand increases so does frequency, waiting times dwindle and the boarding/alighting effect becomes increasingly important. Accordingly, the elasticity of frequency rises from $\eta_F = 1/2$ (square root principle) for $N = 0$ to $\eta_F = 1$ (asymptotic linearity) for $N \to +\infty$. Meanwhile, the elasticity of vehicle size decreases from $\eta_S = 1/2$ to $\eta_S = 0$ (asymptotic constancy). In the limiting case $\delta = 0$ (fixed dwelling time), the elasticities are equal and constant: $\eta_F = \eta_S = 1/2$. We find again the result of Mohring (1972) that optimal frequency follows the square root principle.

As demand keeps increasing, the time required for boarding and alighting also increases, causing between-vehicle congestion in the congested regime and reducing the maximal feasible frequency. Frequency decreases as a result (Figure 1), with $\eta_F < 0$. To compensate for the decrease in frequency, the transit authority strongly increases vehicle capacity, with $\eta_S > 1$ (supralinearity).

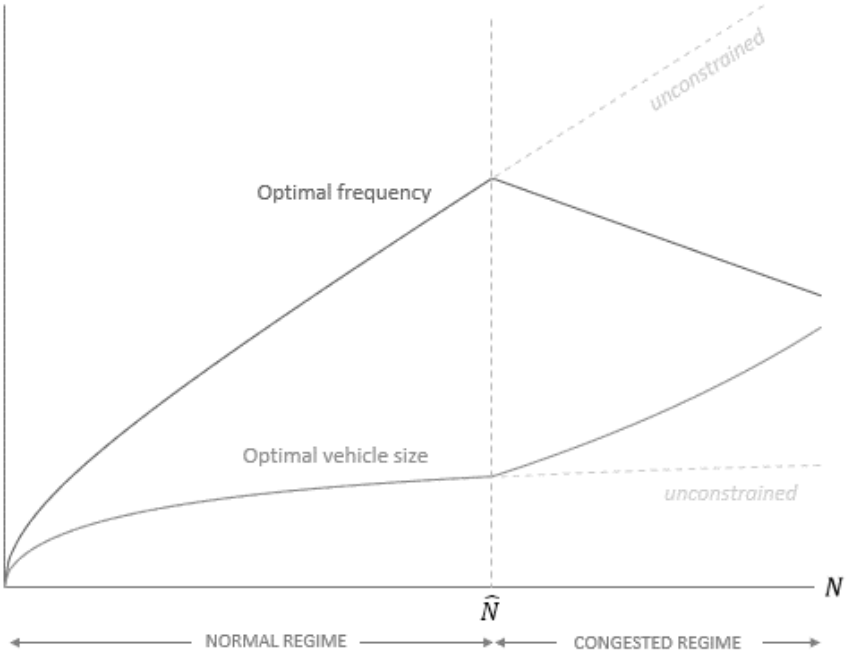


***Figure 1**: Optimal frequency and vehicle size*

3.2. Pricing

Consider the monopolistic case. The profit maximization problem writes: $\max_N N\, G(N) - SC^*(N)$, with $SC^*(N) = SC(F^*(N), s^*(N), N)$. The first-order condition is: $G(N^e) = MSC^*(N^e) - N^e G'(N^e)$. At equilibrium, the marginal user benefit $G(N^e)$ equals the marginal social cost $MSC^*(N^e)$ plus the usual mark-up term $-N^e G'(N^e)$. As $G(N) = A - BN$, the FOC simplifies to: $A - 2BN^e = MSC^*(N^e)$. Solving this equation provides the equilibrium demand level N^e. Finally, the equilibrium fare is:

$$\tau^e = MSC^*(N^e) - N^e G'(N^e) - C_U(F^*(N^e), s^*(N^e), N^e). \tag{15}$$

[11] If vehicle size is exogenous, crowding also causes optimal frequency not to follow the square root principle (see Eq.(33) in the proof of Proposition 1).

Consider now that the transit authority maximizes social welfare. The first-order condition becomes: $G(N^*) = MSC^*(N^*)$.This is the standard result that the marginal user benefit equals the marginal social cost at optimum. We assume in the remainder of the section that this equation is a sufficient condition for optimality.[12] The optimal fare is finally:

$$\tau^* = MSC^*(N^e) - C_U(F^*(N^*), s^*(N^*), N^*). \quad (16)$$

The equilibrium and optimal solutions are characterized by the same service provision rules (provided in Table 1) yet different demand levels. The optimal demand solves $A - BN^* = MSC^*(N^*)$, whereas the equilibrium demand solves $A - 2BN^e = MSC^*(N^e)$, hence Proposition 4.

Proposition 4

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

Vehicle size is also always lower at equilibrium as a result. If optimal demand is low (resp. high), frequency is lower (resp. greater) at equilibrium than at optimum.

Monopolistic behavior involves raising the fare thus reducing demand relatively to the social optimum in order to maximize profit. Facing lower demand, the monopolist also opts for smaller vehicle sizes. In the normal regime, as optimal frequency increases with demand, the monopolist undersupplies frequency. In the congested regime, the opposite occurs, however, as excessive demand leads to congestion between vehicles and overcrowding. The monopolist is able to (and finds it profitable to) oversupply frequency relatively to the social optimum.

4. Economies of scale

4.1. Economies of scale: short run, medium run and long run

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

$$ASC = \frac{\alpha_W}{2F} + \alpha_V d\left(\frac{1}{v_F} + \delta\frac{N}{F}\right) + \alpha_C d\frac{N}{sF} + c_K\frac{sF}{N} + \frac{c_O}{v_F}\frac{F}{N} + c_O\delta. \quad (17)$$

In the short run (fixed frequency and vehicle size), production costs are subject to scale economies (the average production cost C_{TA}/N decreases with the number of trips produced *N*), while user costs are characterized by diseconomies of scale (due to the crowding and boarding/alighting externalities). In the medium run (fixed vehicle size), unit costs are homogeneous of degree 0 with respect to the couple (N, F), except for the waiting cost which decreases with *F*, hence a source of scale economies.

[12] The FOC $G(N^*) = MSC^*(N^*)$ is not a sufficient condition for optimality. The marginal social cost $MSC^*(N)$ is a convex function of N, decreasing from +∞ to its minimum $MSC^*(\widehat{N}) > 0$ on $]0, \widehat{N}]$, then increasing back to +∞ on $[\widehat{N}, N_{max}[$ (Lemma 1 in Appendix). As the inverse demand $GC(N) = A - BN$ is an affine, decreasing function of N, the FOC actually admits either zero, one or two solutions depending on the parameter values (Lemma 2 in Appendix). Moreover, even if $G(N) = MSC^*(N)$ admits one (or two) solution, this solution may only be a local optimum, and the corner solution$N = 0$ may yield a better outcome. Using Lemma 3 (in Appendix), we show that there exists $A_0 > 0$ so that $\forall A \geq A_0$, the equation $G(N) = MSC^*(N)$ admits two solutions, the second one being the actual global optimum.

In the long run, the transit authority can choose to adjust both frequency and vehicle size. Again, user costs present economies of scale with respect to frequency (due to waiting), but diseconomies of scale with respect to vehicle size (as raising vehicle size fails to address the boarding/alighting externality). Conversely, production costs present economies of scale with respect to vehicle size as operating costs (e.g. driver costs) are not affected by it. In addition to raising frequency, expanding vehicle capacity therefore represents another possible source of economies of scale in the long run, up to the tradeoff between economies in operating costs on the one hand and losses in dwelling costs on the other hand. Finally, if demand is too strong, the technical constraint (4) causes frequency to decline, which likely represents a source of diseconomies of scale. Let $i \in \{s; m; l\}$ characterize the horizon considered (short-run *s*, medium-run *m*, long-run *l*). The outcome of these various effects and tradeoffs is provided by Proposition 5.[13]

Proposition 5

In all three time horizons, the provision of the public transit service is subject to economies of scale if $N < N_i$, and to diseconomies of scale if $N > N_i$, where the various N_i solve:

$$\left(\alpha_V \delta d + \alpha_C \frac{d}{\mathrm{s}}\right) {N_s}^2 = \left(c_K s + \frac{c_O}{v_F}\right) F^2, \tag{18}$$

$$\left(\alpha_V \delta d + \alpha_C \frac{d}{\mathrm{s}} + \alpha_W \frac{\delta d_M}{2}\right) {N_m}^2 = \left(c_K s + \frac{c_O}{v_F}\right) {F_0}^2 (1 - \delta d_M N_m)^2, \tag{19}$$

$$\left(\alpha_V \delta d + \alpha_W \frac{\delta d_M}{2}\right) {N_l}^2 = \frac{c_O}{v_F} {F_0}^2 (1 - \delta d_M N_l)^2. \tag{20}$$

In the short run, the marginal user causes the average social cost to decrease at first by splitting fixed production costs between more users. As demand increases, the negative externality - imposing higher in-vehicle-travel time and greater crowding costs to other passengers – eventually prevails, however.

In Mohring's model (1972) the system is always characterized by scale economies in the long run, which stem from the waiting time effect. Despite introducing variable boarding/alighting time and crowding in our model, we find the exact same result in the normal regime: economies of scale are always in order and entirely derive from the waiting time effect, with $dASC^*/dN = -\alpha_W/2F^*N < 0$. This is true both in the medium run (fixed vehicle size) and in the long run (adjustable frequency and vehicle size). As demand increases, the system eventually enters the congested regime (for $N > \widehat{N}$). While economies of scale do persist at first, beyond a second threshold ($N > N_m$ in the medium run and $N > N_l$ in the long run), operations degrade to such an extent that diseconomies of scale occur ($dASC^*/dN > 0$). For convenience, we will refer to this situation characterized by scale diseconomies ($N > N_m$ in the medium run, $N > N_l$ in the long run) as the "hypercongested regime".

[13] Because the equilibrium and optimal solutions are characterized by the same provision rules regarding frequency and vehicle size, the reduced social cost functions $SC^e(N) = SC(F^e(N), s^e(N), N)$ and $SC^*(N) = SC(F^*(N), s^*(N), N)$ are equal: $SC^e(N) = SC^*(N)$. The discussion that follows therefore applies to both cases.

In practice, the optimal (N^*) and equilibrium (N^e) demand levels can be in either regime - normal, congested, hypercongested - depending on the values of the demand parameters *A* and *B* (Figure 2). Here recall that $MSC = ASC + N\ dASC/dN$. The threshold between economies and diseconomies of scale, noted N_l in the long run and defined by $dASC^*/dN(N_l) = 0$, is therefore also the point where the marginal social cost and average social cost curves intersect, hence: $MSC^*(N_l) = ASC^*(N_l)$. Similarly, the limit $\widehat{N}$ between the normal and the congested regime graphically corresponds to the kink in the curve $MSC^*(N)$, where its derivative is discontinuous and shifts from being negative to being positive – meaning that $MSC^*(N)$ is minimized at $N = \widehat{N}$ (Lemma 1).

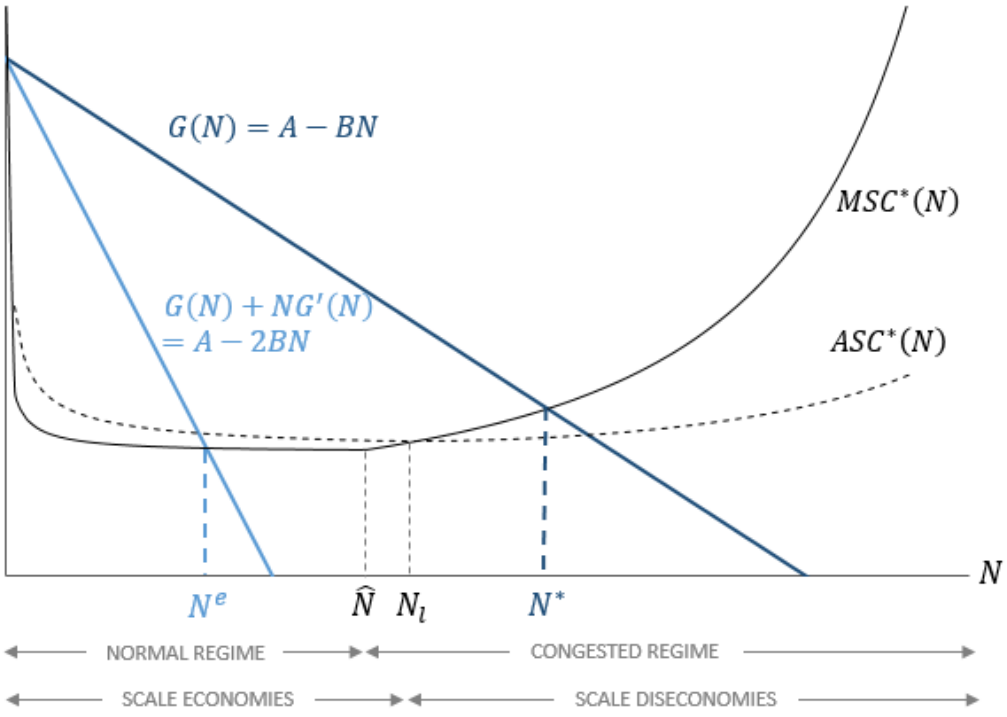


Figure 2: *Optimal and equilibrium demand levels*

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

4.2. Sources of long-run economies of scale

In order to better understand the sources of economies of scale, we break down the long run average social cost by cost item, and study the effect of an increase in demand N. In the normal regime, operating and waiting time costs generate economies of scales, while travel time costs are subject to diseconomies of scale (Table 2). The former effects prevail at first, hence economies of scale overall. In the congested regime, a marginal increase in demand causes frequency to decrease. The waiting time cost per trip increases as a result, so that user costs are characterized by diseconomies of scale. Scale economies associated to operating costs keep prevailing at first, but eventually for $N > N_l$ the negative externalities overweigh the positive ones and the system falls into diseconomies of scale.

***Table 2**: Sources of long-run economies/diseconomies of scale*

	$N \leq \widehat{N}$	$\widehat{N} \leq N < N_l$	$N_l \leq N < N_{max}$
Regime	Normal	Congested	Hypercongested
Waiting time cost / trip	-	+	+
Travel time cost / trip	+	+	+
Crowding cost / trip	=	=	=
Capital cost / trip	=	=	=
Operating cost / trip	-	-	-
Social cost / trip (ASC^*)	-	-	+

Note: a sign + (resp. -) indicates that the corresponding cost per trip increases (resp. decreases) with N, implying diseconomies (resp. economies) of scale. A sign = is used to indicate constancy (no scale economies/diseconomies).

The threshold N_l plays a key role by marking the frontier between congested versus hypercongested regime and economies versus diseconomies of scale. For reminder N_l is defined by:

$$\left(\alpha_V \delta d + \alpha_W \frac{\delta d_M}{2}\right) {N_l}^2 = \frac{c_O}{v_F} {F_0}^2 (1 - \delta d_M N_l)^2.$$

Following the same reasoning as for $\widehat{N}$ (see 3.1 and Figure A.1), one can show that:

- upgrading the transportation technology (decreasing the boarding-alighting time δ and/or the minimum safe headway H_0) improves economies of scale (greater value of N_l);
- increasing interstation distance d_M increases the risk of diseconomies of scale (as it leads to more people at each station and thus longer boarding and alighting times);
- being always characterized by economies of scale (fixed cost effect), greater operating costs (greater parameter c_O) imply greater economies of scale overall (greater value of N_l);
- other parameters (vehicle speed v_F, demand parameters α_W, α_V and d) increase the risk of diseconomies of scale inasmuch as they push the transit authority to raise frequency.

4.3. Implications for funding

We study the impact of (dis) economies of scale on the self-financing of the rail service at the long-run social optimum. From $MSC = ASC + N\ dASC/dN$ and $ASC = C_U + C_{TA}/N$, we can rewrite (16) as the standard first-best pricing rule (Small and Verhoef, 2007, eq (4.44)):

$$\tau^* = \frac{C_{TA}(F^*(N^*), s^*(N^*), N^*)}{N^*} + N^* \frac{dASC^*}{dN}(N^*). \qquad (21)$$

Corollary of Proposition 5

At optimum, the transit service is subsidized if $N^* < N_l$, and self-financed if $N^* \geq N_l$.

Let $\pi^* = C_{TA}/N^* - \tau^*$ be the (long-run) optimal subsidy per trip. In the normal regime, the optimal subsidy equals the average waiting cost, as in Mohring (1972): $\pi^* = \alpha_W/2F^*$. In the congested regime, the optimal subsidy is more complex, with: $\pi^* = (\alpha_W/2 + \alpha_V d/d_M)\delta d_M F_0/F^{*2} - c_O/v_F \times F_0/N^{*2}$. If $\widehat{N} < N^* < N_l$, π^* is positive and the service is subsidized. If $N^* > N_l$, π^* becomes negative due to the Pigouvian principle, in which case the service is (more than) self-financed (Figure 3).

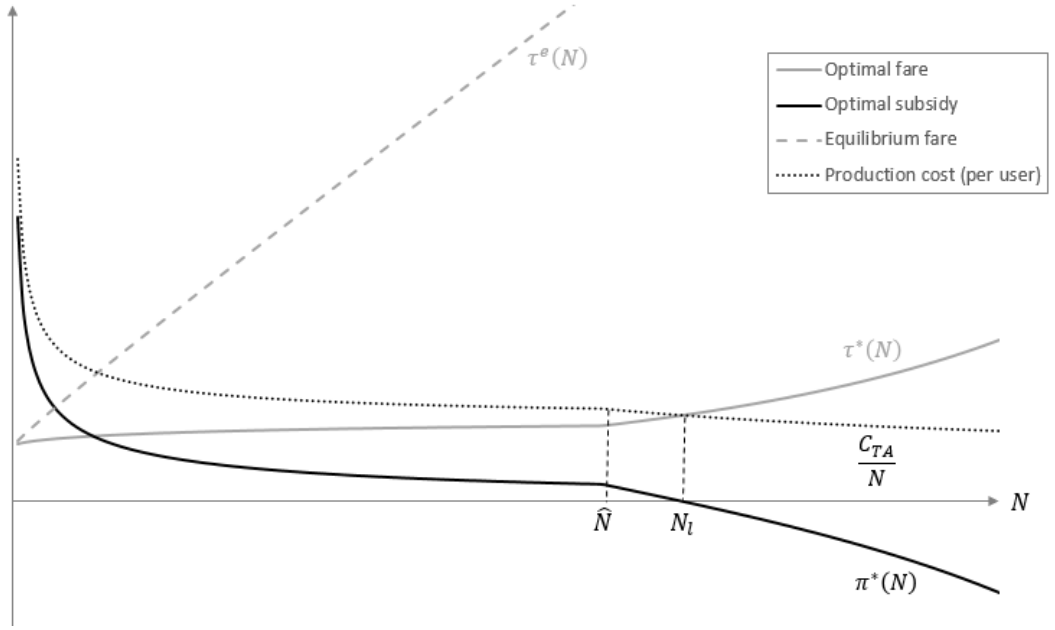


Figure 3: *Optimal (and equilibrium) fare and optimal subsidy*

Note: the threshold value $\widehat{N}$ between the normal and congested regimes corresponds to the kink in the various functions, while the threshold value N_l between economies/diseconomies of scale corresponds to the point where $\tau^*(N)$ and C_{TA}/N intersect.

5. Market distortions

5.1. Marginal cost of public funds

First-best pricing may involve subsidizing the service (Corollary of Proposition 5) and therefore levying taxes in order to cover the public deficit. Such taxes have distortionary effects on the economy, which are commonly captured in the literature by considering a marginal cost of public funds (MCPF), noted μ with $\mu > 0$ (Small and Verhoef, 2007).

The transit authority now maximizes $\int_0^N G(n)dn - SC(F,s,N) - \mu(C_{TA}(F,s,N) - N\tau)$, which is the previous social welfare function minus the cost of funding the deficit of the transit service $C_{TA}(F,s,N) - N\tau$ through (distortionary) taxes. Using the equilibrium condition $G(N) = C_U(F,s,N) + \tau$, the maximization problem rewrites as:

$$\max_{F,s,N} \frac{1}{1+\mu}\int_0^N G(n)dn + \frac{\mu}{1+\mu} N\, G(N) - SC(F,s,N). \quad (22)$$
$$s.t.\ F \leq F_0(1-\delta d_M N)$$

The social welfare maximization problem with a MCPF is analogous to the social welfare and profit maximization problems, the gross benefit term being a weighted average of the two previous ones. Consequently, the optimal provision rules for service frequency and vehicle capacity remain the same. Writing the optimal solution $N^*(\mu)$ as a function of μ, the FOC with respect to demand becomes: $G\big(N^*(\mu)\big) + \mu/(1+\mu)\, N^*(\mu) G'\big(N^*(\mu)\big) = MSC^*\big(N^*(\mu)\big)$.[14] The optimal fare under MCPF is thus:

$$\tau^*(\mu) = \frac{C_{TA}}{N^*(\mu)} + N^*(\mu)\frac{dASC^*}{dN}\big(N^*(\mu)\big) - \frac{\mu}{1+\mu} N^*(\mu) G'\big(N^*(\mu)\big)\,. \tag{23}$$

Compared to (16), the new fare includes an additional term $-\mu/(1+\mu)N^*(\mu)G'\big(N^*(\mu)\big) > 0$ which represents the transit authority incentive to raise the fare in order to reduce the deficit.

Proposition 6

As the marginal cost of public funds μ increases, the optimal fare increases while demand decreases.

As the cost of public money increases, the transit authority raises the fare in order to reduce the deficit (or increase the benefit), as expected. Regarding demand, using $G(N) = A - BN$, the FOC can be rewritten as: $A - \big(1 + \mu/(1+\mu)\big)BN^*(\mu) = MSC^*\big(N^*(\mu)\big)$. It is analogous to the FOC at optimum $A - BN^* = MSC^*(N^*)$ except that the demand parameter B is replaced by $\big(1+\mu/(1+\mu)\big)B$, corresponding to a steeper inverse demand curve. From Figure 2, it is then clear that as μ increases, the optimal demand $N^*(\mu)$ decreases (until converging toward the monopolistic solution as $\mu \to \infty$). From Proposition 3, it follows that the optimal vehicle size decreases with μ. If $N^*(0)$ is in the normal regime, the optimal frequency always decreases with μ. If $N^*(0)$ is in the congested regime however, the optimal frequency increases with μ at first, then decreases.

Considering the marginal cost of public funds leads the transit authority to raise the fare – as a monopolist would do – resulting in a lower demand at optimum. The effect on the optimal provision of service quality is less trivial. The transit authority always reduces vehicle capacity, but may leverage the lower demand to increase frequency in the congested regime, as the cost of adding trains is then more than compensated by the consumer surplus due to the improvement in service quality.

5.2. 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 transit congestion by considering that individuals can choose between two modes: private car (C) and public transit (PT). The number of users are noted N_C and N_{PT}, with $N = N_C + N_{PT}$ the total volume of demand.

To simplify matters, we consider a linear specification for the cost function of car users:

$$C_U^C(N_C) = \beta \frac{N_C}{K}, \tag{24}$$

[14] In the limiting case $\mu = 0$, the FOC degenerates to the FOC obtained at optimum, meaning that $N^*(0) = N^*$ and that our notation is consistent. If $\mu \to +\infty$, the FOC converges this time toward the monopolistic FOC, implying $N^*(\mu) \to N^e$.

where β is the value of travel time by car, and road travel time equals the flow of users N_C divided by (a normalized measure of) road capacity K.[15] Road capacity is fixed, so that the social cost for the road system writes: $SC_C(N_C) = N_C\ C_U^C(N_C) = \beta {N_C}^2/K$. The marginal social cost is: $MSC_C(N_C) = 2\beta N_C/K$. It is linear, and strictly increases for $N_C \in [0, +\infty[$ from 0 to $+\infty$.

Consider first the first-best social welfare maximization problem:

$$\max_{s,F,N_{PT},N_C} \int_0^{N_C+N_{PT}} G(n)dn - SC_C(N_C) - SC_{PT}(F,s,N_{PT}). \qquad (25)$$
$$s.t. \begin{cases} N_C \geq 0 \\ N_{PT} \geq 0 \\ F \leq F_0(1-\delta d_M N_{PT}) \end{cases}$$

From (25), it is straightforward to show that car competition does not change the optimal provision rules for frequency and vehicle size, only the optimal level of demand as stated by Proposition 7.

Proposition 7

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

$$\begin{aligned} &G(N^*) = MSC_C(N^*), N_C^* = N^*,\ N_{PT}^* = 0 && \text{if } N^* \leq N_{PT>0} \\ &G(N^*) = MSC_C(N_C^*) = MSC_{PT}^*(N_{PT}^*) && \text{if } N^* \geq N_{PT>0} \end{aligned} \qquad (26)$$

As the total demand N^* increases, three different patterns successively 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.[16] 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 for too low a demand, it is then more efficient to use only the car (see Figure A.2). 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 optimality is for the social cost of providing public transport to N_{PT}^* users $SC_{PT}^*(N_{PT}^*)$ (green area in Figure A.2) 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 A.2). While this condition is not satisfied at first,[17] eventually it becomes optimal to provide public transit. This causes the number of transit users to jump from 0 to $N_{PT}^* > 0$,

[15] This corresponds e.g. to the equilibrium user cost of a road bottleneck model (Arnott et al., 1993).

[16] Note that because $MSC_C(N^* - N_{PT}) = 2\beta(N^* - N_{PT})/K$ is an affine function of N_{PT}, the problem of solving $MSC_C(N^* - N_{PT}) = MSC_{PT}^*(N_{PT}^*)$ is equivalent to solving the optimal demand equation $G(N^*) = MSC^*(N^*)$ in the standard optimal case without the road mode (see subsection 3.2).

[17] Consider the case where the curves $MSC_C(N^* - N_{PT})$ and $MSC_{PT}^*(N_{PT})$ are tangent and only intersect at N_O. At $N_{PT} = N_O$ the marginal social costs are equated across modes: $MSC_{PT}^*(N_O) = MSC_C(N^* - N_O)$. Yet, the green area $SC_{PT}^*(N_O)$ is strictly larger than the blue quadrilateral $SC_C(N^*) - SC_C(N^* - N_O)$, meaning that it is less costly to transport all N^* users by car, rather than $N^* - N_O$ users by car and N_O users by public transit.

while the number of car users N_C^* 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.

Regarding pricing, the first-best solution is:

$$\tau_C^* = N_C^* \frac{dASC_C}{dN}(N_C^*), \tag{27}$$

$$\tau_{PT}^* = \frac{C_{TA}(F^*(N_{PT}^*), s^*(N_{PT}^*), N_{PT}^*)}{N_{PT}^*} + N_{PT}^* \frac{dASC_{PT}^*}{dN}(N_{PT}^*). \tag{28}$$

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 5) still apply. Because it reduces public transportation demand,[18] 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 $N_C^* \times dASC_C/\,dN$ (operating costs are not considered for this mode).

Consider now the second-best case in which car travel is not taxed ($\tau_C = 0$), so that car users only incur the private cost $C_U^C(N_C)$. The second-best solution is characterized by the following system:

$$\begin{aligned} &G(N^{**}) = C_U^C(N^{**}), N_C^{**} = N^{**},\ N_{PT}^{**} = 0 && \text{if } N^{**} \le N'_{PT>0} \\ &G(N^{**}) = C_U^C(N_C^{**}) = MSC_{PT}^*(N_{PT}^{**}) - \frac{-G'(N^{**})}{C'^C_U(N_C^{**}) - G'(N^{**})}\left(MSC_C(N_C^{**}) - C_U^C(N_C^{**})\right) && \text{if } N^{**} > N'_{PT>0} \end{aligned} \tag{29}$$

$$\tau_C^{**} = 0, \tag{30}$$

$$\tau_{PT}^{**} = \frac{C_{TA}}{N_{PT}^{**}} + N_{PT}^{**} \frac{dASC_{PT}^*}{dN}(N_{PT}^{**}) - \frac{-G'(N^{**})}{C'^C_U(N_C^{**}) - G'(N^{**})}\left(MSC_C(N_C^{**}) - C_U^C(N_C^{**})\right). \tag{31}$$

Compared to the first-best optimum, the 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 the fact that car travel is underpriced. The ratio $-G'/\left(C'^C_U - G'\right)$ being equal to $B/(B + \beta/K) < 1$, the transit fare reduction is less than the implicit car subsidy $MSC_C - C_U^C$ corresponding to the absence of road pricing. If demand is too strong, diseconomies of scale in public transit (captured by the term $N_{PT}^{**}\, dASC_{PT}^*/dN$) exceed this discount, however, resulting in a negative net subsidy: $\tau_{PT}^{**} > C_{TA}/N_{PT}^{**}$.

The above results are summarized in Figure 4. 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

[18] From (26), 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 $MSC_{PT}^*(N_{PT}^*) < G(N_{PT}^*)$. From Figure 2, it is then clear that the solution N_{PT}^* is lower than the solution N^* without the car mode.

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. The impact on the average social cost remains very limited, however.[19]

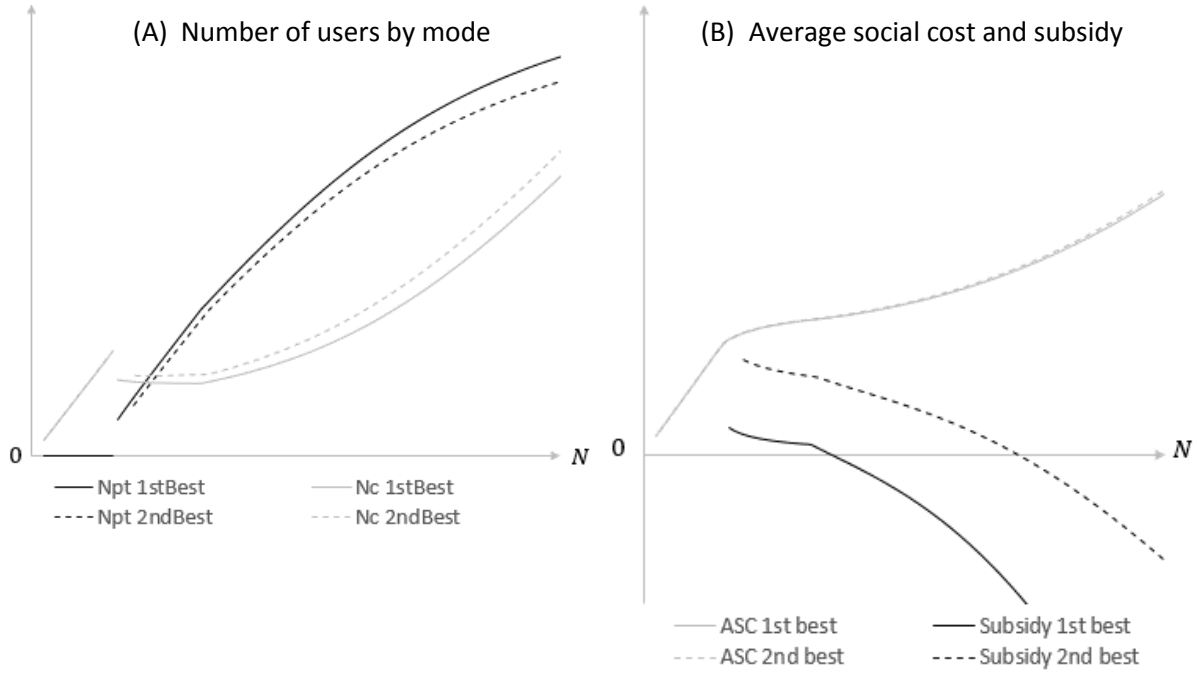


***Figure 4**: Influence of car competition on mode choice (A), the average social cost and subsidies (B)*

6. Data

In order to investigate empirically the existence of scale diseconomies 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.3), 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 culminate at King's Cross St Pancras station. 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.

[19] Because second-best pricing leads to a greater total travel demand than under first-best pricing ($N^{**} > N^{*}$), the negative impact on social welfare (not illustrated here) is substantially more pronounced, meaning that car taxation does matter.

Table 3 reports key figures of the Piccadilly line for the morning peak (7am - 10am) and hyperpeak (8am - 9am) periods, and the corresponding peak direction (westbound).[20] The central section corresponds to the busiest part of the line, which extends from Wood Green to Russell Square station. The Piccadilly line is characterized by an average service quality for a metro line, with a capacity of 684 users per vehicle and an average frequency of 22 trains/h during the morning peak. The frequency is even 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.[21] This results from a substantially stronger demand in the central section - almost four times the average boarding rate (per km) of the 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 3 : 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
Boardings (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). The parameter values are reported in Table 4 and grouped according to three parameter categories: technical, demand, and cost. The following paragraphs briefly discuss the parameter values and how they were computed, including data sources. More extensive information regarding data is available in Online Appendix A.

[20] In 2017, trip direction on the Piccadilly line during the morning peak period (7am – 10am) was split as follows: 56% westbound, 44% eastbound (RODS 2017).

[21] On the busiest interstation (Kings Cross – Russell square), the average load factor even exceeds 100% during the hyperpeak.

Table 4: Parameter values

	Parameter	Value	*Source*
Technical			
d_m	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)	0.43	*Authors' estimate from TfL validation and supply datasets*
Demand			
A	Maximum WTP (£)	14.12	*Authors' estimate from RODS 2017*
B	Slope of WTP (£/user.km^{-1}.h^{-1})	- 0.0040	*Authors' estimate from RODS 2017*
d	Mean trip length (km)	6.07	*Authors' estimate from RODS 2017*
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)*
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)*
μ	Marginal cost of public funds	0.3	*Kleven & Kreiner (2006)*

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 by the largest observed frequency per every 100 tap-in in trains/h) against observed demand (measured by the per km validation rate), using 2013 and 2014 data collected for each day on a hourly basis. More specifically, we estimate the structural equation $F = F_0(1 - \delta d_M N)$ for the congested regime (Figure 5), which allows us to retrieve $H_0 = {F_0}^{-1}$ and δ (as the value of d_M is known). The minimum safe headway estimate is H_0 = 111.8 s, which corresponds to a maximum frequency of 32 trains/h, while the marginal dwelling time estimate is 0.42 s per additional user.[22]

[22] 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-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.

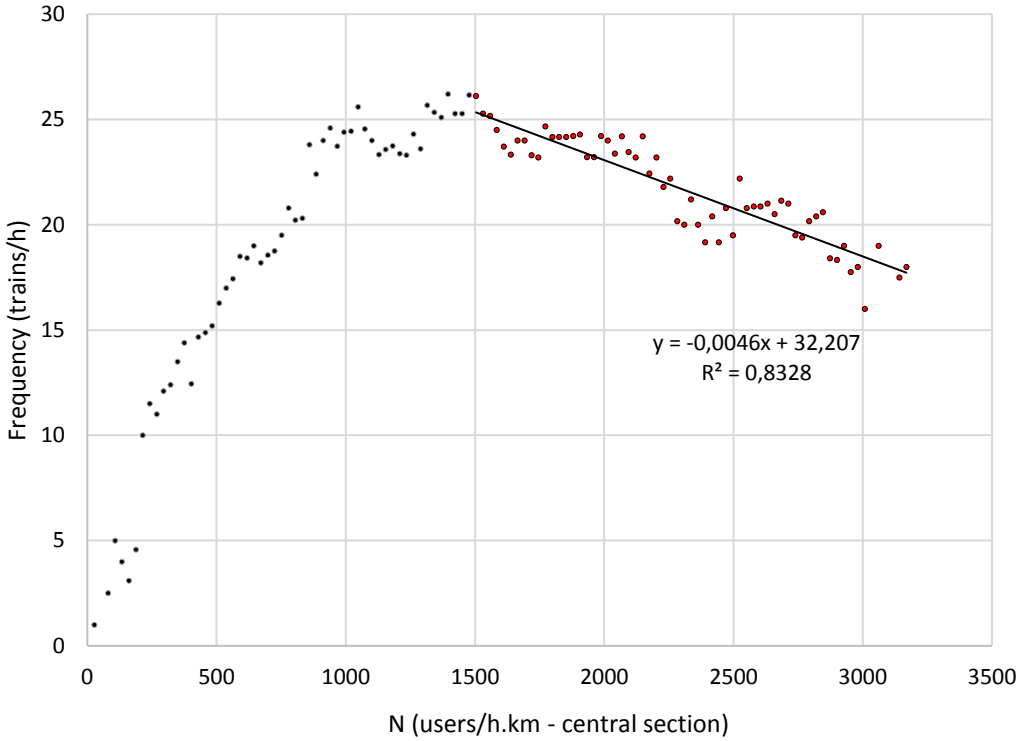


Figure 5: Relationship between real supply and observed demand levels

Now moving to the demand parameters, the linear demand function is estimated by crossing the observed demand level with the generalized price, assuming a generalized price elasticity of -0.75.[23] 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).

7. Empirical results

7.1. Medium run

We first present the medium-run solution - keeping vehicle size fixed - for the three considered provision regimes: monopoly, optimum, and MCPF. The MCPF solution being an intermediate between the monopolistic and optimal solutions, the discussion focuses on the latter two cases.

[23] 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.

As expected, transit fares are greater under monopoly than at optimum (Table 5), 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.[24] All three components of the user cost are greater at optimum than under monopoly: a lower frequency implies greater waiting costs and, combined with a stronger demand, longer boarding and alighting times and greater crowding levels. Conversely, the greater demand plus a lower frequency causes average operating costs to be lower at optimum than under monopoly. All in all, the fare is set above the average operating cost in all three provision regimes and in the observed situation, 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.

Moving to the crux of the paper, we find substantial diseconomies of scale for both the optimum and MCPF cases and the observed situation (as already implied by the large negative subsidy regarding the optimum). In all three cases, the strong level of demand causes overcrowding and congested train operations, which ultimately result in diseconomies of scale. Moderate scale economies persist under monopoly as the lower demand allows for normal train operations.

The breakdown of the optimal subsidy shows that strong scale diseconomies on the demand side are partly offset by supply-side scale economies (Table 6). The crowding effect is largely accountable for the negative subsidy, representing more than two thirds of the (negative) overall user externality. The Mohring effect (waiting time externality) is on the other hand negligible due to the high frequency, contrasting with its preponderance in the theoretical literature. To better understand how and to what extent the overcrowding effect underlies our results, we relax the maximum frequency constraint (4) and compute the optimal subsidy. 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 are markedly different depending on whether one considers the maximum frequency constraint (4) or not. The crowding and travel time externalities in particular become preponderant as frequency declines due to a too strong demand and may no longer be neglected. Regarding supply-side externalities, 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.

[24] 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.

Table 5: Fare, scale economies and welfare estimates (medium run)

	Monopoly	**Optimum**	**MCPF**	**Observed**
Patronage (users/km)	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 (£)	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 (£/user)	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 (£)	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%
Scale economies (£)	0.21	-1.71	-1.08	-1.49
Social welfare (£)	10 140	12 258	12 059	12 081
Average social welfare (£/user)	7.73	5.90	6.53	6.04

Table 6: Breakdown of the optimal subsidy (medium run)

	Between-vehicle congestion	
	with	without
Optimal subsidy (£)	-1.71	0.14
coming from		
waiting externality	-0.10	0.13
travel time externality	-0.70	-0.02
crowding externality	-1.91	-0.04
capital cost externality	0.45	0.03
operating cost externality	0.54	0.04

7.2. 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 vehicle capacities 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. As a matter of fact, whatever the provision regime, the transit agency adjusts vehicle size in the long run in order to reach a same constant target load factor (Proposition 2), here 40%, causing vehicle capacity to be much larger at optimum in response to the stronger demand.

The increase in vehicle size 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 level allows for normal operations, with 21.6 trains/h (against 25.4 trains/h in the medium run). By adjusting vehicle size, the transit agency is able to operate the line more efficiently. In the long run the service provision is therefore characterized by significantly lower scale diseconomies at optimum (with or without MCPF), and by slightly larger scale economies under monopoly, all contributing to the lower transit fares in the long run than in the medium run.

Table 7: Fare, scale economies and welfare estimates (long run)

	Monopoly	**Optimum**	**MCPF**	**Observed**
Patronage N (users/km)	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 (£)	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 (£/user)	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 (£)	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%
Scale economies (£)	0.25	-0.61	-0.23	-1.49
Social welfare (£)	10 225	13 313	12 960	12 081
Average social welfare (£/user)	7.74	5.49	6.30	6.04

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 larger one (with -0.94 £ per additional 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 short run.

Table 8: Breakdown of the optimal subsidy (long run)

	Between-vehicle congestion	
	with	without
Optimal subsidy (£)	-0.61	0.15
coming from		
waiting externality	-0.13	0.13
travel time externality	-0.94	-0.05
crowding externality	0	0
capital cost externality	0	0
operating cost externality	0.47	0.07

7.3. Off-peak

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

Table 9: Fare, scale economies and welfare estimates (off-peak, medium run)

	Monopoly	**Optimum**	**MCPF**	**Observed**
Patronage N (users/km)	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 (£)	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*
TA cost (£/user)	1.01	0.90	0.93	0.99
vehicle capital costs	*0*	*0*	*0*	*0*
operating costs	*1.01*	*0.90*	*0.93*	*0.99*
Price (£)	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%
Scale economies (£)	0.40	0.22	0.27	0.28
Social welfare (£)	1 874	2 567	2 469	2 422
Average social welfare (£/user)	4.35	2.96	3.51	3.62

As demand is lower during the off-peak period, trains operate normally in all cases considered (Table 9). This leads to standard results from the literature, such as a greater frequency and lower user costs at optimum than under monopoly, and to (moderate) scale economies 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 10). 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.

Table 10: Breakdown of the optimal subsidy (off-peak, medium run)

	Between-vehicle congestion	
	with	without
Optimal subsidy (£)	0.22	0.22
coming from		
waiting externality	0.19	0.19
travel time externality	-0.02	-0.02
crowding externality	-0.06	-0.06
capital cost externality	0	0
operating cost externality	0.11	0.11

7.4. 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 5) to 21.8 trains/h (Table 11), while the load factor increases from 81% to 92%. The line is subject to even greater scale diseconomies as a result, as a marginal user bears on other users an additional 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 11: Fare, scale economies and welfare estimates of the NTfL program

	Baseline	**NTfL**
Patronage N (users/km)	2 269	2 642
Regime	*Congested*	*Congested*
Frequency (trains/h)	21.8	25.1
Vehicle capacity (users)	684	889
User cost (£)	3.41	2.93
waiting	*0.24*	*0.21*
in-vehicle travel time	*1.64*	*1.54*
crowding	*1.53*	*1.18*
TA cost (£/user)	0.79	0.75
vehicle capital costs	*0.28*	*0.28*
operating costs	*0.51*	*0.47*
Price (£)	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%
Scale economies (£)	-2.29	-1.56
Social welfare (£)	13 844	15 842
Average social welfare (£/user)	6.10	6.00

8. 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.

Without between-vehicle congestion, our model would always predict economies of scale in the medium run and long run, whatever the level of demand. This contrasts with Tirachini et al. (2010) who find that crowding eventually results in scale diseconomies in the medium run (fixed vehicle size). The difference in our results - linked to the use of a quadratic function of the vehicle load factor for the crowding cost as opposed to a linear function in our case - underlines the significant influence of model specification in determining the final balance between economies and diseconomies of scale. Assuming stronger negative externalities - as in Tirachini et al. (2010) regarding crowding - would increase diseconomies of scale, whereas assuming greater supply-side economies of scale or stronger positive user externalities – e.g. with regard to the Mohring effect - would yield the opposite. Similarly, the choice of a simple linear inverse demand function in our model 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 show that congestion between vehicles is a major source of diseconomies of scale for heavily used transit lines – as 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 - that may not be neglected and should be addressed by appropriate policies (such as pricing or technological upgrades).

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 (between lines) and in time (between peak and off-peak). 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). Network adjustments could alleviate congestion in the very long run by designing alternate lines for the most popular OD pairs, 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 new infrastructure solutions, however.

Among the 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. Preliminary computations show that making boarding/alighting time a function of vehicle size delays the occurrence of overcrowding (as bigger vehicles handle boardings and alightings 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 *non-so uncommon* cases, the answer should be "*yes*", and if heavily crowded urban transit system remain subsided, it should not be motivated by the usual rationales (economies of scale and underpricing of car travel).

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

Adler, M.W., van Ommeren, J.N., 2016. Does public transit reduce car travel externalities? Quasi-natural experiments' evidence from transit strikes. Journal of Urban Economics 92, 106–119. https://doi.org/10.1016/j.jue.2016.01.001

Anderson, M.L., 2014. Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion. American Economic Review 104, 2763–2796. https://doi.org/10.1257/aer.104.9.2763

Arnott, R., de Palma, A., Lindsey, R., 1993. A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. American Economic Review 83, 161–179.

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

De Borger, B., Proost, S., 2012. A political economy model of road pricing. Journal of Urban Economics 71, 79–92. https://doi.org/10.1016/j.jue.2011.08.002

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

Farsi, M., Fetz, A., Filippini, M., 2007. Economies of Scale and Scope in Local Public Transportation. Journal of Transport Economics and Policy 41, 345–361.

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.

Glaister, S., Lewis, D., 1978. An integrated fares policy for transport in London. Journal of Public Economics 9, 341–355. https://doi.org/10.1016/0047-2727(78)90015-4

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

Jansson, J.O., 1980. A Simple Bus Line Model for Optimisation of Service Frequency and Bus Size. Journal of Transport Economics and Policy 14, 53–80.

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

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.

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

Nash, C., Sansom, T., Still, B., 2001. Modifying transport prices to internalise externalities: evidence from European case studies. Regional Science and Urban Economics, Evaluating Policies to Reduce Transportation Air Pollution 31, 413–431. https://doi.org/10.1016/S0166-0462(01)00059-X

Nelson, P., Baglino, A., Harrington, W., Safirova, E., Lipman, A., 2007. Transit in Washington, DC: Current benefits and optimal level of provision. Journal of Urban Economics, Essays in Honor of Kenneth A. Small 62, 231–251. https://doi.org/10.1016/j.jue.2007.02.001

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.

Ripplinger, D.G., Bitzan, J.D., 2018. The cost structure of transit in small urban and rural U.S. communities. Transportation Research Part A: Policy and Practice 117, 176–189. https://doi.org/10.1016/j.tra.2018.08.021

Savage, I., 2010. The dynamics of fare and frequency choice in urban transit. Transportation Research Part A: Policy and Practice 44, 815–829. https://doi.org/10.1016/j.tra.2010.08.002

Savage, I., Small, K.A., 2010. A Comment on "Subsidisation of Urban Public Transport and the Mohring Effect." Journal of Transport Economics and Policy 44, 373–380.

Silva, H.E., Verhoef, E.T., 2013. Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges. Journal of Public Economics 106, 1–13. https://doi.org/10.1016/j.jpubeco.2013.06.007

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., 2010. 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

van Reeven, P., 2008. Subsidisation of Urban Public Transport and the Mohring Effect. Journal of Transport Economics and Policy 42, 349–359.

Vickrey, W., 1980. Optimal transit subsidy policy. Transportation 9, 389–409. https://doi.org/10.1007/BF00177700

Viton, P.A., 1992. Consolidations of scale and scope in urban transit. Regional Science and Urban Economics 22, 25–49. https://doi.org/10.1016/0166-0462(92)90024-U

Whelan, G., Crockett, J., 2009. An investigation of the willingness to pay to reduce rail overcrowding, in: 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 – Additional figures

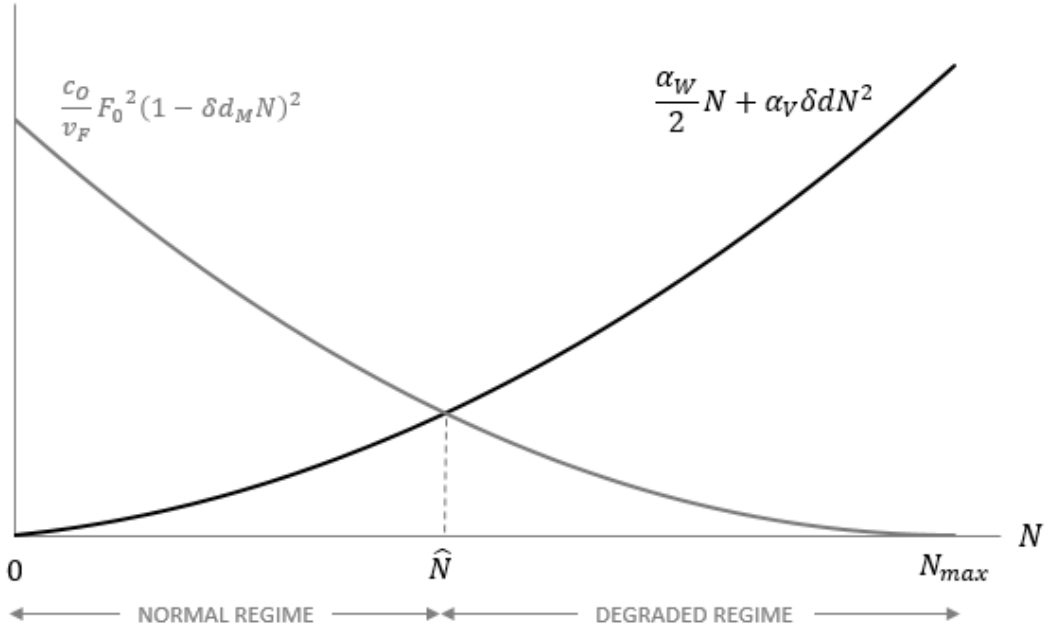
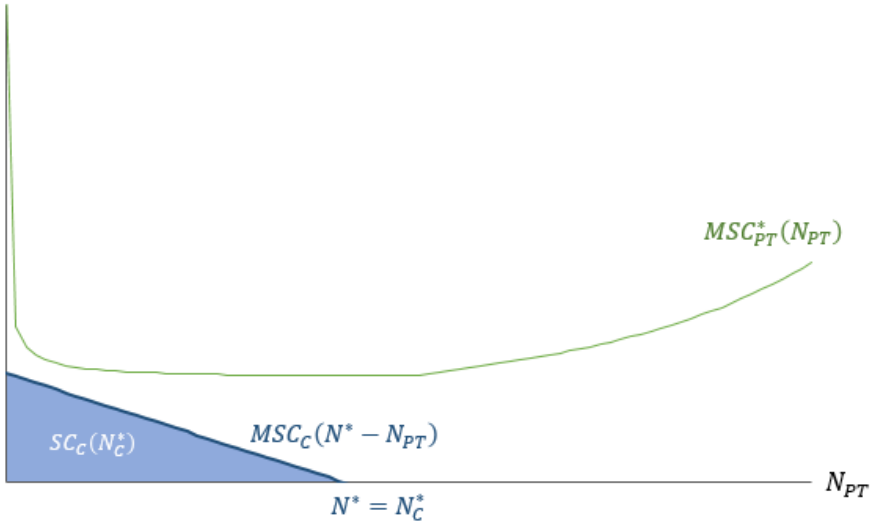
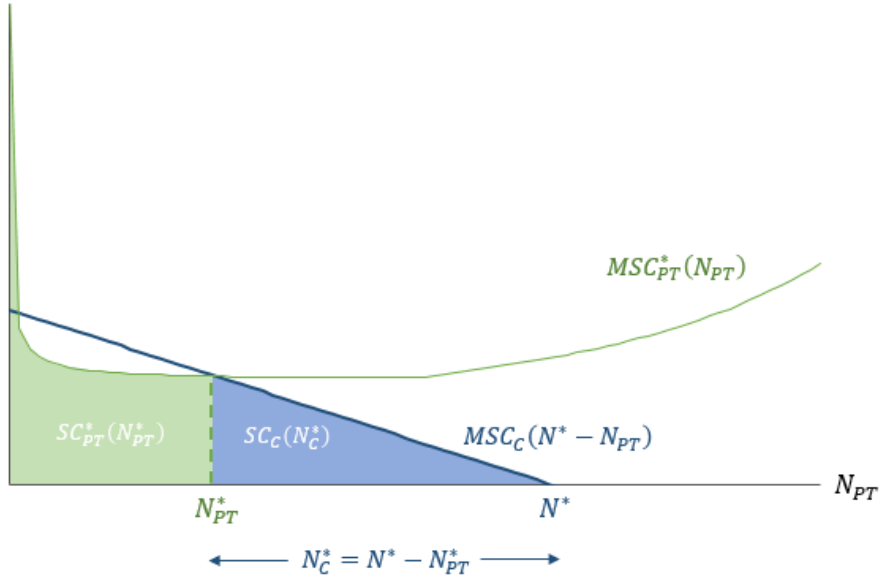


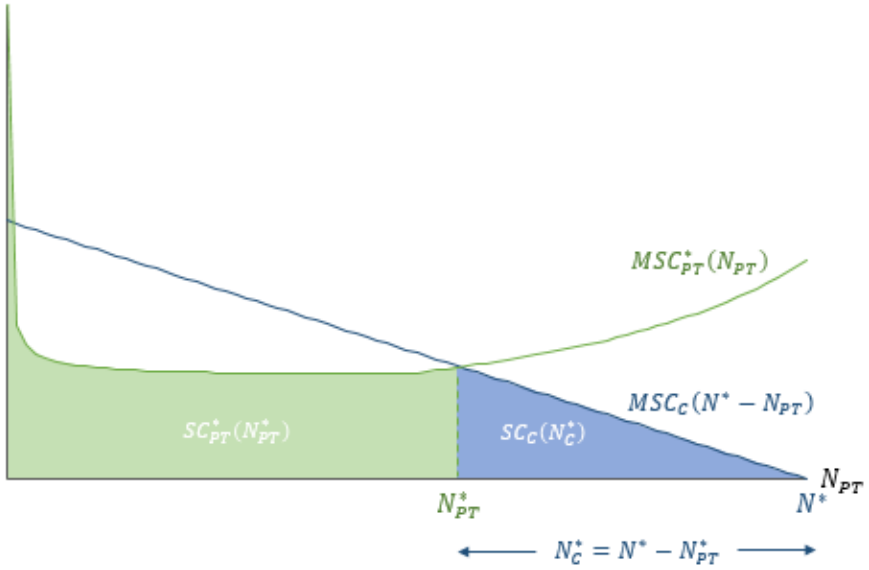
***Figure A.1**: Limit $\widehat{N}$ between the normal and congested regimes*



(a) Low total demand



(b) Medium total demand



(c) High total demand

***Figure A.2**: Optimal demand levels for car and for public transit*

Note: because $MSC_C(0) = 0$, the curve $MSC_C(N^* - N_{PT})$ always intersects the abscissae axis at $N_{PT} = N^*$.

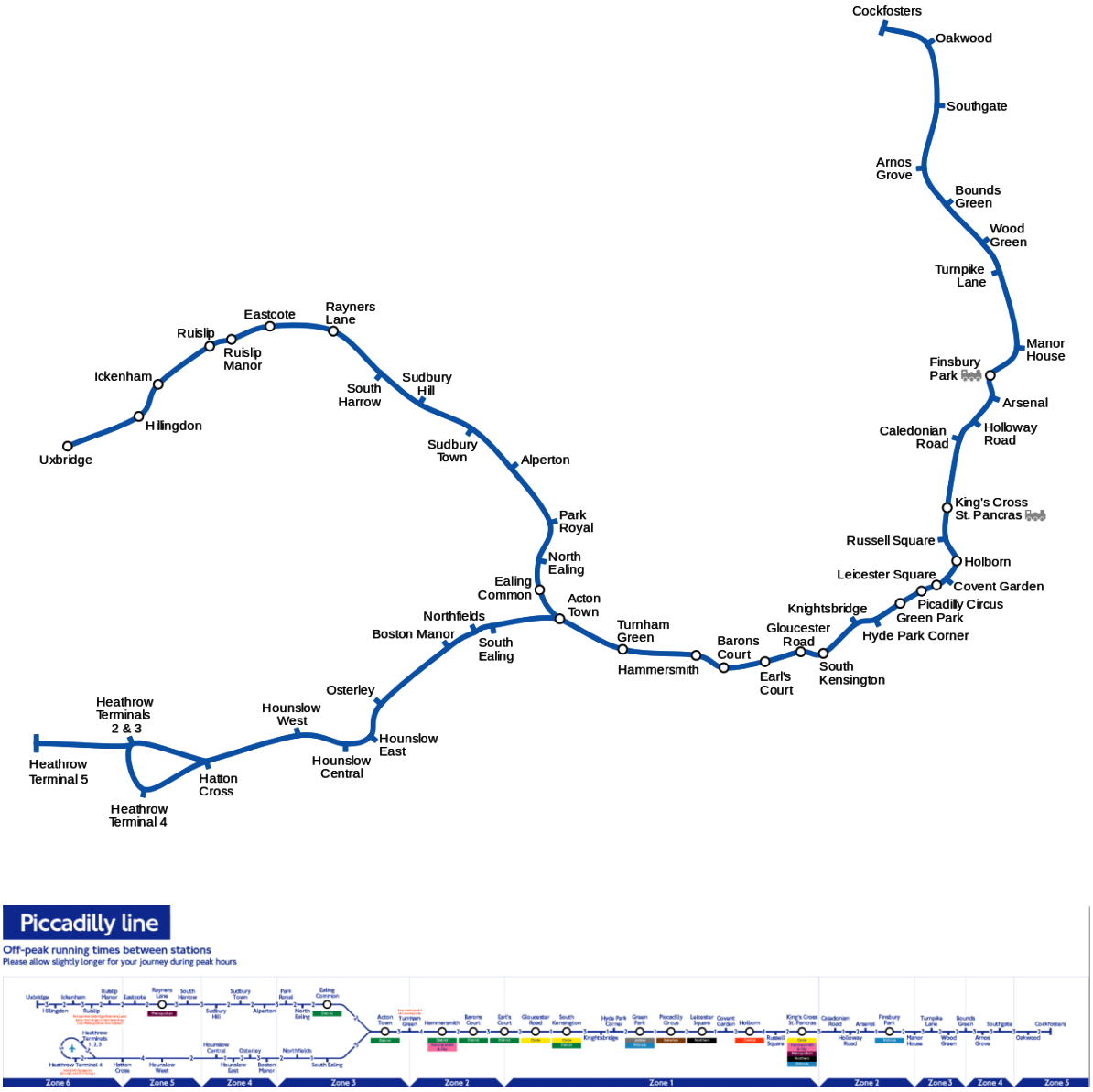


Figure A.3: Map of the Piccadilly line

Appendix B - Proofs

Proposition 1 (optimal frequency)

The Lagrangian corresponding to the constrained minimization problem writes:

$$\mathcal{L} = SC(F, s, N) + \lambda\big(F - F_0(1 - \delta d_M N)\big).$$

The first-order condition (FOC) relative to service frequency is:

$$\frac{\partial \mathcal{L}}{\partial F} = \frac{\partial SC}{\partial F} + \lambda = -\frac{N}{F^{*2}}\left(\frac{\alpha_W}{2} + \alpha_V d\delta N + \alpha_C d\frac{N}{s^*}\right) + c_K s^* + \frac{c_O}{v_F} + \lambda = 0.$$

Assume first that the constraint (4) is not binding: $\lambda = 0$. Raising frequency decreases waiting costs, in-vehicle costs (through shorter boarding-alighting times) and crowding costs (through lower vehicle loads). On the other hand, it raises operating costs and capital costs as it involves an increase in both vehicle-hours and vehicle-kilometers. The FOC can be rewritten as:

$$c_O\left(\frac{F^*}{v_F} + \delta N\right) + c_K s^* F = N\left(\frac{\alpha_W}{2F^*} + \alpha_V \frac{\delta dN}{F^*} + \alpha_C \frac{dN}{s^* F^*}\right) + c_O \delta N\,, \tag{32}$$

which is Proposition 1. Rearranging the above equality yields:

$$F^* = \sqrt{\frac{1/2\ \alpha_W N + dN^2(\alpha_V \delta + \alpha_C / s^*)}{c_K s^* + c_O / v_F}}. \tag{33}$$

If vehicle size is exogenous (medium-run adjustment), the optimal frequency F^* does not follow the square root principle because of: 1) variable alighting-boarding times and 2) in-vehicle crowding.

Next assume that the constraint (4) is binding. The optimal frequency equals the maximal feasible one: $F^* = F_0(1 - \delta d_M N)$. □

Proposition 2 (vehicle size)

The FOC relative to vehicle size *s* is:

$$\frac{\partial \mathcal{L}}{\partial s} = \frac{\partial SC}{\partial s} = -\alpha_C d\frac{N^2}{s^{*2} F^*} + c_K F^* = 0,$$

$$\Leftrightarrow \qquad N\alpha_C \frac{dN}{s^* F^*} = c_K s^* F^*. \tag{34}$$

Increasing vehicle size reduces crowding costs, but increases capital costs. At optimum, the two are equal, hence Proposition 2.

We can then rewrite (34) as $s^* F^* = \sqrt{\alpha_C d\ / c_K} N$, so that the optimal load factor is:

$$l_C{}^* = \frac{dN}{s^* F^*} = \sqrt{\frac{c_K d}{\alpha_C}}.$$

It is constant and independent from *N* and *F*. □

Table 1 (optimal frequency and vehicle size)

Consider first the normal regime. Combining the two FOC yields:

$$\frac{\alpha_C d N^2}{c_K s^{*2}} = \frac{1/2\ \alpha_W N + dN^2(\alpha_V \delta + \alpha_C / s^*)}{c_K s^* + c_O / v_F},$$

$$\Leftrightarrow\ \ \alpha_C d s^* N + \alpha_C d \frac{1}{c_K}\frac{c_O}{v_F} N = s^{*2}(\alpha_W/2 + \alpha_V \delta d N) + \alpha_C d s^* N.$$

The last equation simplifies to:

$$s^* = \sqrt{\frac{c_O / v_F}{c_K}\frac{\alpha_C d N}{\frac{\alpha_W}{2} + \alpha_V \delta d N}}.$$

Using the FOC relative to vehicle size, we also get:

$$F^* = \sqrt{\frac{v_F}{c_O}\left(\frac{\alpha_W}{2} N + \alpha_V \delta d N^2\right)}.$$

We now compute the optimal solution in the congested regime. The optimal frequency is dictated by technological constraints:

$$F^* = F_0(1 - \delta d_M N).$$

From the FOC relative to vehicle size, the optimal vehicle size is thus:

$$s^* = \sqrt{\frac{\alpha_C d}{c_K}\frac{N}{F_0(1 - \delta d_M N)}}.$$

Before determining the boundary between the normal regime and the congested regime, let us first note that given the maximum frequency condition (4), we must have $\delta d_M N < 1$; otherwise there are so much users that people do not have the time board, resulting in a negative maximum frequency. We note $N_{max} = 1/\delta d_M$ this upper limit for N.

The normal regime corresponds to the case $F^* < F_0(1 - \delta d_M N)$. In light of the above, this is equivalent to $N < \widehat{N}$, where $\widehat{N}$ is the first positive solution of:

$$\frac{\alpha_W}{2}\widehat{N} + \alpha_V \delta d \widehat{N}^2 = \frac{c_O}{v_F} {F_0}^2 \left(1 - \delta d_M \widehat{N}\right)^2.$$

Because the RHS zeroes in $N = N_{max}$, the above equation has two positive solutions, a first one lower than N_{max}, then a second one greater than N_{max}. As the second one leads to a negative frequency ($N > N_{max} \Rightarrow F_0(1 - \delta d_M N) < 0$), $\widehat{N}$ is the first positive (and only acceptable) solution. □

Proposition 3 (elasticities)

The elasticities in the normal regime are given by:

$$\eta_S = \frac{\partial s^* / \partial N}{s^* / N} = \frac{1}{2} - \frac{\alpha_V \delta N}{2\left(\frac{\alpha_W}{2d} + \alpha_V \delta N\right)} = \frac{1}{2}\left(1 - \frac{\alpha_V \delta d N}{\frac{\alpha_W}{2} + \alpha_V \delta d N}\right) \Rightarrow 0 \leq \eta_S \leq 1/2,$$

$$\eta_F = \frac{\partial F^*/\partial N}{F^*/N} = \frac{1}{2}\frac{\frac{\alpha_W}{2}N + 2\alpha_V \delta d N^2}{\frac{\alpha_W}{2}N + \alpha_V \delta d N^2} = \frac{1}{2}\left(1 + \frac{\alpha_V \delta d N}{\alpha_W/2 + \alpha_V \delta d N}\right) \Rightarrow 1/2 \leq \eta_F \leq 1.$$

As demand increases, the public transport authority increases frequency more and more, because the extra boarding time effect plays a more important role (by comparison with the waiting time effect). In the limiting case $N \to +\infty$, we have:

$$\lim s^* = \sqrt{\frac{c_O/v_F}{c_K}\frac{\alpha_C}{\alpha_V \delta}} \text{ and } F^* \sim \sqrt{\frac{\alpha_V \delta d}{c_O/v_F}}N.$$

Vehicle size converges toward a constant (the elasticity converges toward 0) and frequency increases linearly in *N* (the elasticity converges toward 1). These limits should be interpreted as asymptotic behaviors for large values of *N*, as for $N > \widehat{N}$ the system switches to the congested regime.

In the congested regime, the optimal frequency $F^* = F_0(1 - \delta d_M N)$ decreases with *N*, hence $\eta_F < 0$. The demand elasticity of vehicle size is:

$$\eta_S = 1 + \frac{F_0 \delta d_M N}{F_0(1 - \delta d_M N)} = \frac{1}{1 - N/N_{max}} > 1 \; for \; N \in \left[\widehat{N}, N_{max}\right].$$

Vehicle size increases supra-linearly with *N.* □

Proposition 4

The optimal demand solves:

$$G(N^*) = A - BN^* = MSC^*(N^*).$$

As discussed in Footnote 12, this equation admits solutions (either one or two) only if *A* is large enough. As we assume throughout this paper that this is indeed the case, Lemma 3 states that N^* is the second (positive) solution to $A - BN = MSC^*(N)$. Noting N^0 the first solution, we have $A - BN > MSC^*(N)$ if $N \in]N^0, N^*[$ and $A - BN < MSC^*(N)$ if $N \in [0, N^0[\cup]N^*, +\infty[$.

The (monopoly) equilibrium demand solves:

$$G(N^e) + N^e G'(N^e) = A - 2BN^e = MSC^*(N^e).$$

This implies $A - BN^e = MSC^*(N^e) + BN^e > MSC^*(N^e)$. Based on the above, $N^e \in]N^0, N^*[$, meaning that $N^e < N^*$, which is the first part of Proposition 4.

From Proposition 3, it follows that $s^e < s^*$. In the normal regime $F^e < F^*$, while in the congested regime $F^e > F^*$, which is the second part of Proposition 4.

Proposition 5

Consider first the short-run optimum. The derivative of the average social cost is:

$$\frac{dASC}{dN}(N) = \left(\alpha_V \delta + \frac{\alpha_C}{s}\right)\frac{d}{F} - \left(c_K s + \frac{c_O}{v_F}\right)\frac{F}{N^2}.$$

The system is characterized by economies of scale ($dASC/dN > 0$) if demand is lower than N_s, and diseconomies of scale ($dASC/dN < 0$) if it is greater than N_s, where N_s zeroes the above equation.

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:

$$\frac{dASC^*}{dN}(N) = \left(\alpha_V \delta + \frac{\alpha_C}{s}\right)\frac{d}{F^*} - \left(c_K s + \frac{c_O}{v_F}\right)\frac{F^*}{N^2}.$$

Using the FOC relative to frequency, the formula simplifies to:

$$\frac{dASC^*}{dN}(N) = -\frac{\alpha_W}{2F^*N} < 0.$$

There are always economies of scale in the normal regime.

In the congested regime, the constraint is active so that we can no longer use the envelope theorem. Instead, we directly compute the long-run average social cost, which is:

$$ASC^*(N) = \frac{\alpha_V d}{v_F} + \frac{\alpha_W/2 + (\alpha_V d\delta + \alpha_C d/s)N}{F_0(1-\delta d_M N)} + \left(c_K s + \frac{c_O}{v_F}\right)\frac{F_0(1-\delta d_M N)}{N} + c_O\delta.$$

Differentiating the previous expression with respect to *N* yields:

$$\frac{dASC^*}{dN}(N) = \frac{\alpha_V d\delta + \alpha_C d/s + \delta d_M \alpha_W/2}{F_0(1-\delta d_M N)^2} - F_0\frac{c_K s + c_O/v_F}{N^2}.$$

In the congested regime, the system is therefore characterized by economies of scale if $N < N_m$, and diseconomies of scale if $N > N_m$, where $N_m$is the (first) positive solution of:

$$(\alpha_V d\delta + \alpha_C d/s + \alpha_W \delta d_M/2)N^2 = {F_0}^2(c_K s + c_O/v_F)(1-\delta d_M N)^2. \quad (35)$$

Consider finally the long-run optimum. In the normal regime, using the same method as previously, we again find that $dASC^*/dN = -\alpha_W/2F^*N < 0$ and that there are always economies of the scale.

In the congested regime, the long-run average social cost is now given by:

$$ASC^*(N) = \alpha_V\frac{d}{v_F}\left(1 - \frac{v_F}{d_M F_0}\right) + 2\sqrt{d\alpha_C c_K} + c_O\delta\left(1 - \frac{d_M F_0}{v_F}\right) + \frac{c_O}{v_F}\frac{F_0}{N} + \frac{\frac{\alpha_W}{2} + \alpha_V\frac{d}{d_M}}{F_0(1-\delta d_M N)}.$$

The derivative of the average social cost is:

$$\frac{dASC^*}{dN}(N) = \left(\frac{\alpha_W}{2} + \alpha_V\frac{d}{d_M}\right)\frac{\delta d_M}{F_0(1-\delta d_M N)^2} - \frac{c_O}{v_F}\frac{F_0}{N^2}. \quad (36)$$

(36) is negative for $N < N_l$, and positive for $N > N_l$, where N_l is the (first) positive solution of:

$$\frac{\alpha_W}{2}\delta d_M \bar{N}^2 + \alpha_V \delta d \bar{N}^2 = \frac{c_O}{v_F}{F_0}^2(1-\delta d_M \bar{N})^2. \quad (37)$$

Proposition 6

We first show that optimal demand decreases with μ. The FOC with respect to demand writes:

$$A - \left(1 + \frac{\mu}{1+\mu}\right) BN^*(\mu) = MSC^*\big(N^*(\mu)\big).$$

The RHS $MSC^*(N)$ is a convex function strictly decreasing then strictly increasing with *N* (Lemma 1). Conversely, the LHS is an affine, strictly decreasing function of *N*. Because its slope decreases with μ, it is straightforward to show by adapting Lemma 2 that $N(\mu)$ also decreases with μ.

Next to show that the optimal fare $\tau(\mu)$ increases with μ, we use the user equilibrium condition $GC(N) = C_U(F^*(N), s^*(N), N) + \tau$. The curves $GC(N)$ and $C_U(F^*(N), s^*(N), N)$ are given. Because $GC(N)$ decreases with N, if $N^*(\mu)$ decreases with μ¸ it follows that $\tau^*(\mu)$ must increase with μ.

Proposition 7

As in Section 3, we first solve for the optimal frequency and vehicle size as functions of N_{PT}. We can therefore rewrite the maximization problem (25)

as:

$$\max_{s,F,N_{PT},N_C} \int_0^{N_C+N_{PT}} GC(n)dn - SC_C(N_C) - SC^*_{PT}(N_{PT}). \tag{38}$$

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

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 $MSC_C(0) = 0 < MSC^*_{PT}(N^*)$, keeping the total number of users N^* constant, the social welfare can be increased by switching an infinitesimal quantity ε>0 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: $GC(N^*) = MSC_C(N^*_C) = MSC^*_{PT}(N^*_{PT})$. Using $MSC_C(N_C) = 2\beta N_C/K$ and $N^*_C + N^*_{PT} = N^*$, this implies that N^*_{PT} satisfies:

$$\frac{2\beta(N^* - N^*_{PT})}{K} = MSC^*_{PT}(N^*_{PT}). \tag{39}$$

The LHS is an affine, strictly decreasing function of N_{PT}. MSC^*_{PT} is a strictly convex function of N_{PT}, decreasing from +∞ to a minimum $MSC^*_{PT}(\widehat{N}) > 0$ on $[0, \widehat{N}]$, then increasing again to +∞ on $[\widehat{N}, N_{max}]$. Using Lemma 2, there exists $N_0 > 0$ so that (39) admits no solution if $N^* < N_0$, exactly one solution if $N^* = N_0$ (when the two curves become tangent), then exactly two solutions if $N^* > N_0$ (cf**. Figure A.2** for a graphical intuition).

If $N^* \leq N_0$ we are therefore always in case 2, meaning it is more efficient to transport all users by car. If $N^* > N_0$, whether one is in case 2 or in case 3 depends on whether the social cost of providing public transport to N_{PT}^* users $SC_{PT}^*(N_{PT}^*)$ is lower than that of transporting the same amount of users by car, which is $SC_C(N^*) - SC_C(N^* - N_{PT}^*)$. Using Lemma 3 with $f(x) = MSC_C(N^* - x)$ and $g(x) = MSC_{PT}^*(x)$, there exists $N_{PT>0} > N_0$ so that if $N^* \leq N_{PT>0}$ then the corner solution $N_{PT}^* = 0$ and $N_C^* = N^*$is the global optimum, while if $N^* \geq N_{PT>0}$, then the interior solution N_{PT}^* and $N_C^* = N^* - N_{PT}^*$, where N_{PT}^* is the second solution to Equation (39), is the global optimum.

We now show the second part of Proposition 6. 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 $N_{PT}^* > 0$. Because N_{PT}^* is the second solution to Equation (39), using Lemma 2 we know that N_{PT}^* increases with N^* and converges toward N_{max} as $N^* \to +\infty$. Accordingly, as long as $N_{PT}^* \leq \widehat{N}$ the marginal social cost $MSC_{PT}^*(N_{PT}^*)$ decreases as N^* and N_{PT}^* increase, 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^* as long as $N_{PT}^* \leq \widehat{N}$. If $N_{PT}^* \geq \widehat{N}$ however, then $MSC_{PT}^*(N_{PT}^*)$ increases this time as N^* and N_{PT}^* increase, implying that N_C^* increases again.

Appendix C - Lemmas

Lemma 1

The marginal social cost is a convex function of N, decreasing from +∞ to its minimum $MSC^*(\widehat{N}) > 0$ on $]0, \widehat{N}]$, then increasing back to +∞ on $[\widehat{N}, N_{max}[$.

Proof

In the normal regime ($N \in]0, \widehat{N}]$), the frequency constraint (4) is inactive. Consequently, we can apply the envelope theorem to compute the marginal social cost, which yields:

$$MSC^*(N) = \frac{\partial SC}{\partial N}(F^*(N), s^*(N), N).$$

For reminder, the social cost is:

$$SC(F, s, N) = \frac{\alpha_W}{2}\frac{N}{F} + \alpha_V d\left(\frac{N}{v_F} + \delta\frac{N^2}{F}\right) + \alpha_C d\frac{N^2}{sF} + c_K sF + \frac{c_O}{v_F}F + c_O\delta N.$$

The marginal social cost therefore reads:

$$MSC^*(N) = \frac{\alpha_W}{2F^*(N)} + \alpha_V d\left(\frac{1}{v_F} + \delta\frac{N}{F^*(N)}\right) + \alpha_C d\frac{N}{s^*(N)F^*(N)} + c_O\delta.$$

Using Proposition 1 and Table 1, this simplifies to:

$$MSC^*(N) = \frac{\alpha_V d}{v_F} + \sqrt{\alpha_C c_K d} + c_O\delta + \sqrt{\frac{c_O}{v_F}\left(\frac{\alpha_W}{2}\frac{1}{N} + \alpha_V\delta d\right)}.$$

From the above expression, it is clear that $MSC^*(N)$ strictly decreases on $]0, \widehat{N}]$, and that its limit in $N = 0^+$ is $+\infty$. By differentiating $MSC^*(N)$ twice, the convexity is also straightforward to show.

In the congested regime ($N \in [\widehat{N}, N_{max}[$), the minimum social cost is given by:

$$SC^*(N) = \alpha_V \frac{d}{v_F}\left(1 - \frac{v_F}{d_M F_0}\right)N + 2\sqrt{d\alpha_C c_K}N + c_O\delta\left(1 - \frac{d_M F_0}{v_F}\right)N + \frac{c_O}{v_F}F_0 + \frac{\left(\frac{\alpha_W}{2} + \alpha_V \frac{d}{d_M}\right)N}{F_0(1 - \delta d_M N)}.$$

The marginal social cost is thus:

$$MSC^*(N) = \alpha_V \frac{d}{v_F}\left(1 - \frac{v_F}{d_M F_0}\right) + 2\sqrt{d\alpha_C c_K} + c_O\delta\left(1 - \frac{d_M F_0}{v_F}\right) + \frac{\frac{\alpha_W}{2} + \alpha_V \frac{d}{d_M}}{F_0(1 - \delta d_M N)^2}.$$

Again, from the above expression it is straightforward to show that $MSC^*(N)$ is also convex on $[\widehat{N}, N_{max}[$, strictly increases on this interval and tends toward $+\infty$ as $N \to {N_{max}}^-$.

All in all, this shows that $MSC^*(N)$ is convex on $]0, N_{max}[$, strictly decreases from $+\infty$ to its minimum $MSC^*(\widehat{N}) > 0$ on $]0, \widehat{N}]$, then strictly increases from $MSC^*(\widehat{N})$ to $+\infty$ on $[\widehat{N}, N_{max}[$.

Lemma 2

Let $f(x) = a - bx$, with $a \geq 0$ and $b \geq 0$ be an affine, decreasing function of x.

Let $g(x)$ be a positive, strictly convex function defined on an open interval $]0, x_{max}[$, with the following limits: $\lim_{x \to 0^+} g(x) = \lim_{x \to {x_{max}}^-} g(x) = +\infty$.

There exists $a_0 > 0$ so that the equation $f(x) = g(x)$ admits (exactly) zero solution if $a < a_0$, one solution if $a = a_0$, and two solutions if $a > a_0$. Let $x_1(a, b)$ and $x_2(a, b)$, with $x_1(a, b) < x_2(a, b)$, denote the two solutions in the latter case. Then $x_1(a, b)$ and $x_2(a, b)$ decrease and increase with a, respectively, with $\lim_{a \to +\infty} x_1(a, b) = 0$ and $\lim_{a \to +\infty} x_2(a, b) = x_{max}$.

Proof

For the sake of concision, we only provide a graphical intuition of the first part of the proof. Based on Figure 3, it is clear that as *a* increases:

- at first the curve $f(x) = a - bx$ remains below $g(x)$ so that the curves never intersect;
- at some value of $a = a_0$, $f(x)$ is tangent to $g(x)$: the curves only intersect in one point;
- then $f(x)$ intersects $g(x)$ in exactly two points due to the strict convexity of *g*.

We note $x_1(a, b)$ and $x_2(a, b)$, with $x_1(a, b) < x_2(a, b)$, the two solutions if $a > a_0$. Considering that $\lim_{x \to 0^+} g(x) = \lim_{x \to {x_{max}}^-} g(x) = +\infty$, and that f is bounded on $[0, x_{max}]$, then $f(x) < g(x)$ if $x \epsilon]0, x_1(a, b)[\cup]x_2(a, b), +\infty[$ and $f(x) > g(x)$ if $x \epsilon]x_1(a, b), x_2(a, b)[$.

Next consider $a_2 > a_1 > a_0$. We have $a_2 + bx_1(a_1, b) > a_1 + bx_1(a_1, b) = g\big(x_1(a_1, b)\big)$, meaning that $x_1(a_1, b) > x_1(a_2, b)$ (since $f(x) > g(x) \Longrightarrow x\epsilon]x_1(a, b), x_2(a, b)[$). Similarly we have $x_2(a_1, b) < x_2(a_2, b)$.

Regarding the limits, $f(x) \geq a - bx_{max} \forall x \in [0, x_{max}]$. As $a \to +\infty$, $a - bx_{max}$ also tends toward $+\infty$, meaning that $\lim_{a\to+\infty} g\left(x_1(a, b)\right) = \lim_{a\to+\infty} g\left(x_2(a, b)\right) = +\infty$. Because $x_1(a, b) < x_1(a_0, b)$ and $x_2(a, b) > x_2(a_0, b)$ $\forall a > a_0$ it follows that $\lim_{a\to+\infty} x_1(a, b) = 0$ and $\lim_{a\to+\infty} x_2(a, b) = x_{max}$.

Lemma 3

Let $f(x) = a - bx$, with $a \geq 0$ and $b \geq 0$ be an affine, decreasing function of x.

Let $g(x)$ be a positive, strictly convex function defined on an open interval $]0, x_{max}[$, and integrable over the same interval, with the following limits: $\lim_{x\to0^+} g\,(x) = \lim_{x\to x_{max}^-} g\,(x) = +\infty$.

Consider the maximization problem: $\max_{x\geq0} \int_0^x \big(f(u) - g(u)\big)du$.

There exists $a_1 > a_0$ so that if $a < a_1$, the maximum is reached at $x = 0$ while if $a > a_1$ the maximum is reached at $x = x_2(a, b)$, where a_0 and $x_2(a, b)$ are defined in Lemma 2.

Proof

Let $H(x) = \int_0^x \big(f(u) - g(u)\big)du$ defined for $x \in [0, x_{max}]$. Its derivative is $H'(x) = f(x) - g(x)$.

If $a \leq a_0$, from Lemma 2 $f(x) \leq g(x)$ $\forall x \in]0, x_{max}[$ and $H(x)$ decreases on $[0, x_{max}]$. In this case, it always reaches its maximum for *x*=0.

Next consider $a > a_0$. In this case the function $H(x)$ decreases on $[0, x_1(a, b)]$, then increases on $[x_1(a, b), x_2(a, b)]$, and finally decreases on $[x_2(a, b), x_{max}]$. There are two possibilities: $H(x)$ is maximized either at *x*=0 or at *x*=$x_2(a, b)$. The question is then whether $H(0) \lesseqgtr H\big(x_2(a, b)\big)$. We have:$H\big(x_2(a, b)\big) = \int_0^{x_2(a,b)} \big(f(u) - g(u)\big)du$. The derivative of $H\big(x_2(a, b)\big)$ with respect to a is $x_2(a, b) - \frac{\partial x_2(a,b)}{\partial a}\Big(f\big(x_2(a, b)\big) - g\big(x_2(a, b)\big)\Big) = x_2(a, b) > 0$. It follows that $H\big(x_2(a, b)\big)$ strictly increases with a. Because $H\big(x_2(a_0, b)\big) < 0$, Lemma 3 straightforwardly follows.