Demand Imbalances and Multi-Period Public Transport Supply

Daniel Hörcher* and Daniel J. Graham

Railway and Transport Strategy Centre
Department of Civil Engineering
Imperial College London

Manuscript submitted for Transportation Research Part B
Initial submission: 30 May 2017
Resubmission: 15 December 2017

Abstract
This paper investigates multi-period public transport supply, i.e. networks in which capacity cannot be differentiated between links and time periods facing independent but nonidentical demand conditions. This setting is particularly relevant in public transport, as earlier findings on multi-period road supply cannot be applied when the user cost function, defined as the sum of waiting time and crowding costs, is nonhomogeneous. The presence of temporal, spatial and directional demand imbalances is unavoidable in a public transport network. It is not obvious, however, how the magnitude of demand imbalances may affect its economic and financial performance. We show in a simple back-haul setting with elastic demand, controlling for total willingness to pay in the network, that asymmetries in market size reduce the attainable social surplus of a service, while variety in maximum willingness to pay leads to higher aggregate social surplus and lower subsidy under efficient pricing. The analysis of multi-period supply sheds light on the relationship between urban structure, daily activity patterns, and public transport performance.

Keywords
public transport; transport supply; crowding; capacity optimisation; demand imbalances

*Corresponding author – Electronic address: d.horcher@imperial.ac.uk
1 Introduction

Public transport is supplied by multi-product firms in the sense that services are normally provided along predetermined lines with several stops in both directions. From an economic point of view, each direction of each inter-station section in various time periods can be considered as an individual market with or without demand interactions between them. Due to operational constraints, these markets are often served with the same capacity. As demand for public transport is hardly identical in separate markets, first-best capacity provision can never be feasible in reality. From the operator’s point of view, this constraint translates into the fact that public transport services are subject to demand imbalances, and a second-best capacity has to be determined in a multi-period framework.

The presence of demand fluctuations is hardly questionable. However, their magnitude may differ across a wide range. In this paper, we focus on the simplest case of transport supply under demand imbalances: the back-haul problem, with independent demand curves. We show using a supply optimisation model that the magnitude of demand imbalances can have a crucial impact on the average crowding experience of passengers. Moreover, beyond the optimal capacity, the economic and financial performance of the service is also affected by the differences in ridership in jointly served markets, controlling for the aggregate scale of operations. In Section 3.2 we show that the bigger the deviation in market size, the lower the amount of social surplus that a public operator can achieve, and the more subsidies it will need to cover its losses. By contrast, imbalances in willingness to pay between joint markets reduce the optimal subsidy and leave more aggregate benefits for society. The core message of the paper is that the level of demand asymmetry as an external factor has significant impact on the economic and financial performance of public transport provision.

The magnitude of demand imbalances in the context of road provision is out of the main focus of transport research. Small and Verhoef (2007, Section 5.1.1) derive that the standard self-financing result of Mohring and Harwitz (1962) survives in the multi-period setting as well, assuming (i) constant returns to scale in congestion technology, (ii) neutral scale economies in capacity provision, (iii) perfectly divisible capacity, and (iv) time-varying first-best static congestion pricing where the toll equals to the marginal external congestion cost. The first assumption implies that the user cost function is homogeneous of degree zero, so that Euler’s theorem can be applied to relate the impact of marginal demand and capacity deviations on the user cost of travelling. The empirical literature confirms that the three assumptions on cost functions are not far from reality in road transport, and therefore the transport economics community did not see much potential in further investigating the optimisation of road supply in a multi-period context specifically\(^1\). In public transport,\(^2\)

\(^1\)Exceptions including Bichsel (2001) and Lindsey (2009) focused on second-best scenarios with pricing restrictions and uncertainty.

\(^2\)
however, user costs are far from constant returns to scale due to the well-known Mohring effect (Mohring, 1972). This paper fills in an important gap in the literature with the analysis of multi-period supply optimisation in public transport.

The paper presents two main lines of research. It contributes to the literature of public transport economics with a number of additional theoretical insights. We discuss

- **T1** the cohabitation of Mohring-type waiting time benefits and negative crowding externalities in a public transport model,
- **T2** the application of the Cost Recovery Theorem in the presence of waiting time as well as crowding externalities, and
- **T3** the interplay between frequency and vehicle size provision when demand is unevenly distributed between jointly served markets.

The more policy oriented branch of the paper investigates the effect of the magnitude of demand imbalances on

- **P1** second-best choice of frequency and vehicle size,
- **P2** the resulting peak and off-peak occupancy rates,
- **P3** maximum social surplus that can be reached with second-best supply, considering constant total willingness to pay for the service, and
- **P4** the amount of subsidy which is required to cover the financial deficit under efficient pricing.

The upcoming sections are structured as follows. Section 2 sets the field for subsequent analyses with a baseline supply optimisation model and a discussion of theoretical research questions T1 and T2 in the list above. Then, Section 3, the backbone of the paper, deals with the investigation of second-best supply in the back-haul problem. In particular, Section 3.1 begins with a simple inelastic demand setting which enables us to uncover the mechanics behind theoretical topic T3 above, while Section 3.2 presents core insights on major research objectives P1–P4 in connection with demand imbalances. The most relevant research outcomes are summarised in Table 3. Finally, Section 4 outlines an agenda for future research and Section 5 concludes.

## 2 Fundamentals of public transport supply

In transport economics theory, the main topics of interest in supply optimisation include (i) decision rules for optimal capacity setting, (ii) the determinants of short-run marginal social costs of service usage that form the basis for efficient pricing, and (iii) the degree of self-financing under socially optimal pricing. This section follows the same steps of analysis for the specific case of public transport.
2.1 Earlier literature on public transport capacity

Jara-Díaz and Gschwender (2003) provide a comprehensive review of the evolution of early capacity models. Most of these contributions kept the methodological framework of assuming inelastic demand, constructing a social cost function, and minimising it with respect to the optimal frequency and other supply-side variables.

**Waiting time:** The most common elements of public transport models since Mohring (1972) consider waiting time as a user cost and frequency as a decision variable. These imply scale economies in user costs, as high demand leads to high frequency, low headways, and lower expected waiting time for all users. We further investigate this mechanism in Section 2.3.

**Cycle time:** Several authors model that cycle time (i.e. the running time of vehicles) may be a function of the number of boarding and alighting passengers at intermediate stops through dwell times. This makes the case for a negative consumption externality, because boarding imposes additional travel time cost on passengers already on board. This feature is an important component of capacity models focusing primarily on bus operations\(^2\), e.g. Jansson (1980), Jara-Díaz and Gschwender (2003), Jara-Díaz and Gschwender (2009), Tirachini et al. (2010) and Tirachini (2014).

**Crowding:** Modelling vehicle capacity is another area in which the evolution of the literature can be identified. Most of the early studies expressed vehicle capacity in terms of the maximum number of passengers explicitly (Jansson, 1980, 1984). This approach is convenient from a methodological point of view due to the ease of analytical constrained optimisation, but neglects the user cost of crowding up until the exogenous capacity limit is reached. Empirical evidence shows that passengers may be annoyed even by fellow users sitting on neighbouring seats (Wardman and Murphy, 2015), so user costs may increase under full seat occupancy as well. Oldfield and Bly (1988) assumed that the main impact of high vehicle occupancy on users is that it increases the probability that passengers cannot board the first vehicle, thus lengthening the expected waiting time. Based on demand modelling results, reviewed by Wardman and Whelan (2011) and Li and Hensher (2011), the impact of crowding on the value of in-vehicle travel time can be quantified. Jara-Díaz and Gschwender (2003) extended Jansson (1980) with a crowding dependent linear value of time multiplier. They found that the optimal occupancy rate is independent of demand with neutral scale economies in operational and user costs.

**Daily supply optimisation** with common fleet size has been on the research agenda since Newell (1971), Oldfield and Bly (1988), and Chang and Schonfeld (1991), with no

\(^2\text{Note, however, that travel times of rail services are generally much less sensitive to the number of boarding and alighting passengers than buses with a front door boarding policy. Assuming endogenous train length, dwell times are definitely not linear in the number of boardings, because the optimal number of doors may increase with demand.}\)
specific focus on characterising the pattern of demand. Rietveld (2002) argued that many rail operators are unable to reduce capacity between the morning and afternoon peak, and therefore their sole objective is to meet the highest peak demand. In his setting the marginal social cost of an off-peak trip is basically zero, while in the rush hours the marginal burden is very high, because the incremental capacity often remains in operation throughout the entire day. Guo et al. (2017) show that even if within-day frequency adjustment is possible, changing schedules can be costly from an operational point of view.

The peak load problem was considered in spatial and directional terms as well. Rietveld and Roson (2002) and Rietveld and van Woudenberg (2007) investigated second-best policies in pricing and capacity provision. Rietveld and Roson (2002) found that even under monopolistic behaviour with profit maximising objective, price differentiation between directions can be beneficial from a social welfare perspective. Rietveld and van Woudenberg (2007) compared the welfare loss that uniformed supply-side variables cause in fluctuating demand and revealed that differentiated service frequency would provide significantly more benefits for society than dynamic pricing or adjustable vehicle size. Jansson et al. (2015) builds his model upon the critical section of the main haul (peak direction) in the back-haul problem. He argues that occupancy charges should only apply for those who are on board in the critical section and thus have an impact on fleet size, and boarding and alighting charges should be paid by main haul users only, assuming that schedules have to be identical in the back haul. It is notable that all the above mentioned studies approached the peak load problem with explicit capacity constraints and neglect the external cost of crowding disutilities, thus leaving a gap in the literature.

Our baseline model in Section 2.2 builds on Pels and Verhoef (2007) who selected waiting time and crowding discomfort as the main user cost components. They identified the second-best nature of public transport capacity and derived in their paper’s appendix the optimal frequency and vehicle size rules for multiple markets. However, they did not investigate the impact of demand fluctuations on supply variables. We intend to fill in the highly relevant gap.

2.2 Baseline model

This paper considers a publicly owned transport operator with a welfare maximising objective. The operator has full control over two capacity variables: frequency and vehicle size, where the former is measured as the number of services per hour and the latter is the available floor area inside the vehicle. For the sake of simplicity, sitting and standing are not differentiated, comfort related user costs only depend on the average density of passengers.

For an in-depth analysis of the impact of demand fluctuations on optimal seat provision, the reader is kindly referred to Section 3.2 of Hörcher et al. (2018).
per unit of floor area. All other aspect of capacity provision, including long-term decisions on the infrastructure and other engineering variables, are neglected throughout this study.

We define social welfare as the sum of user benefits ($B$) net of user costs ($C_u = Q \cdot c_u$) and operational costs ($C_o$):

$$SW(F, S, Q) = B - C_u - C_o, \quad (1)$$

where $F$ and $S$ are the frequency and vehicle size set by the operator, and $Q$ is hourly demand. Our next step is to develop a social cost function that captures the main characteristics of public transport operations: we attach importance to density economies in vehicle size\(^4\) and the fact that capacity shortages cause crowding and inconvenience for passengers.

Therefore we merge the operational cost specification of Rietveld et al. (2002) with the crowding multiplier approach of Jara-Díaz and Gschwender (2003), and define the following cost functions:

$$C_o = (v + wS^\delta) \cdot Ft,$$

$$C_u = \alpha \frac{1}{2F} \cdot \underbrace{Q}_{\text{waiting time}} + \beta t \left(1 + \frac{\varphi Q}{FS}\right) \cdot \underbrace{Q}_{\text{in-veh. time & crowding}}. \quad (2)$$

The expected waiting time cost is half of the headway ($0.5F^{-1}$) multiplied by the value of waiting time ($\alpha$), assuming that passengers arrive randomly at the station\(^5\). Furthermore, $t$ is the exogenous travel time, $\beta$ is the value of uncrowded in-vehicle travel time, and the last element of the user cost expression is a crowding-dependent linear travel time multiplier function that reflects the inconvenience of crowding. Here we express the occupancy rate of vehicles with the ratio of demand and capacity, $Q/(FS)$. The user cost of a unit of in-vehicle travel time is linear in the occupancy rate with slope $\varphi$. We assume that travel time ($t$) is independent of the operator’s capacity decisions. Thus, the model is applicable to any transport modes where door capacity is increased proportionally with vehicle size, so boarding and alighting times do not need to be modelled explicitly. Alternatively, one may assume that $\varphi$ in the multiplier function takes account of both crowding disutilities and the impact of excess demand on dwell times.

Operational costs are modelled as the product of total vehicle-hours supplied ($Ft$) and the unit cost of a vehicle-hour, $wS^\delta$, where $\delta$ is the vehicle size elasticity of operational costs. This variable captures the purely technological feature that the average operational costs of a unit of in-vehicle capacity decreases with the size of vehicles. Finally, we add a purely frequency dependent component to the objective function to reflect driver costs, the price of

---

\(^4\)Scale economies may be present in frequency as well, given the fixed cost of infrastructure provision. As fixed costs do not depend on the decision variables of this model, we can safely normalise them to zero without loss of generality.

\(^5\)In the rest of the paper we introduce $a = \alpha/2$ and express the expected waiting time cost as $aF^{-1}$. 
train paths supplied by the infrastructure manager, or the operational cost of a locomotive when applicable, and other expenses.

On the benefit side we introduce \(d(Q)\) as the inverse demand function, i.e. the measure of marginal willingness to pay for the public transport service. In equilibrium, inverse demand equals to the generalised price of travelling, which includes the average user cost \((c_u)\) and the fare \((p)\), so that

\[
d(Q) = c_u(Q, F, S) + p.
\] (3)

It is a fundamental feature of most public transport services that the same capacity has to serve multiple markets subject to varying demand conditions. In order to accommodate demand fluctuations within our model, we assume \(m\) independent markets where fares, represented by vector \(p = (p_1, \ldots, p_m)\), can be differentiated between the markets. We define the following \(\mathcal{L}\) Lagrangian function of the constrained welfare maximisation problem, with \(\lambda = (\lambda_1, \ldots, \lambda_m)\) denoting the vector of Lagrange multipliers of the equilibrium constraint in each market.

\[
\max_{Q(p), F, S, \lambda} \mathcal{L} = \sum_{i=1}^{m} \left[ \int_0^{Q_i} d_i(q) dq - Q_i \left[ aF^{-1} + \beta t_i (1 + \varphi Q_i (FS)^{-1}) \right] \right] - \left( \sum_{i=1}^{m} t_i \right) F (v + wS^\delta) - \sum_{i=1}^{m} \lambda_i \left[ d_i(Q_i) - aF^{-1} - \beta t_i (1 + \varphi Q_i (FS)^{-1}) - p_i \right]
\] (4)

First order conditions with respect to frequency and vehicle size yield the following capacity rules:

\[
aF^{-2} \left( \sum_i Q_i \right) + \beta \varphi S^{-1} F^{-2} \left( \sum_i t_i Q_i^2 \right) = \left( \sum_i t_i \right) \left( v + wS^\delta \right),
\] (5)

\[
\beta \varphi F^{-1} S^{-2} \left( \sum_i t_i Q_i^2 \right) = \left( \sum_i t_i \right) wS^{\delta-1} F.
\] (6)

The optimal frequency equates the marginal benefits of having shorter headways and less crowding due to capacity expansion with the marginal increase in operational cost. The same applies for the optimal vehicle size, where the benefit side is limited to crowding effects only (i.e. there is no waiting time effect).
From the first order conditions of equation (4) with respect to $Q_i$, $p_i$ and $\lambda_i$, we can derive the optimal fare ($p_i$) for market $i$ as well. As the first-best set of optimal fares ensures efficiency on each market, the Lagrange multiplier of the equilibrium constraint drops to zero. The market dependent fare becomes

$$p_i = Q_i \cdot c'(Q_i) = \beta t_i \cdot \frac{Q_i}{FS},$$

which is, unsurprisingly, the marginal external crowding cost imposed on $Q_i$ fellow passengers. This result is in line with Pels and Verhoef (2007) with an explicit specification of marginal external crowding costs. The intuition behind this result is simple when there is no capacity adjustment at all: as the marginal passenger boards the vehicle, the density of crowding increases by $1/(FS)$, which causes disutility for all other travellers. By contrast, when capacity is adjustable, the operator may try to internalise the crowding externality with increased capacity\(^6\). However, equations (5) and (6) ensure that frequency and vehicle size expansion has no welfare effect on the margin when capacity is optimal, so the magnitude of the aggregate social cost of the marginal trip is still equivalent to the theoretical direct crowding externality. In Section 2.3 and 3.1 we investigate how this marginal social cost is split between the Mohring effect, other indirect capacity externalities and operational costs.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Dimension</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q$</td>
<td>Demand</td>
<td>pass/h</td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>Service frequency</td>
<td>1/h</td>
<td></td>
</tr>
<tr>
<td>$S$</td>
<td>Vehicle size</td>
<td>m(^2)</td>
<td></td>
</tr>
<tr>
<td>$t$</td>
<td>In-vehicle travel time</td>
<td>h</td>
<td>0.25 (15min)</td>
</tr>
<tr>
<td>$a$</td>
<td>Half of the value of waiting time</td>
<td>$/h$</td>
<td>15</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Value of uncrowded in-vehicle time</td>
<td>$/h$</td>
<td>20</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Occupancy rate, $\phi = Q/(FS)$</td>
<td>pass/m(^2)</td>
<td></td>
</tr>
<tr>
<td>$\varphi$</td>
<td>Crowding multiplier parameter</td>
<td>(pass/m(^2))(^{-1})</td>
<td>0.15</td>
</tr>
<tr>
<td>$v$</td>
<td>Fixed operational cost per train hour</td>
<td>$/h$</td>
<td>500</td>
</tr>
<tr>
<td>$w$</td>
<td>Variable operational cost per hour per m(^2)</td>
<td>$/(m^2 \cdot h)$</td>
<td>10</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Elasticity of operational costs w.r.t. vehicle size</td>
<td>–</td>
<td>0.8</td>
</tr>
<tr>
<td>$\omega$</td>
<td>Share of peak market in total riderhips</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>$\theta$</td>
<td>Share of peak market in aggregate consumer benefit</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>$A_0$</td>
<td>Maximum willingness to pay at $\theta = 0.5$</td>
<td>$</td>
<td>30</td>
</tr>
<tr>
<td>$M_0$</td>
<td>Market size at $\theta = 0.5$</td>
<td>pass/h</td>
<td>5000</td>
</tr>
</tbody>
</table>

\(^6\)Internalisation in the operator’s context means that some of the externality borne by consumers can be transformed into visible operational costs. This is an alternative policy of internalising the externality through pricing, in which case it is added to the costs borne by the marginal user herself.
Numerical example: First-best capacity

Let us illustrate in a numerical example how capacity variables react to changes in demand. With only one market considered \((m = 1)\), for equilibrium demand level \(Q\) the solution of (4) simplifies to the following cost minimisation problem.

\[
\min_{F, S} TC(F, S, Q) = C_u + C_o = a F^{-1} Q + \beta t \left[ 1 + \varphi Q(FS)^{-1} \right] Q + (v + wS^\delta) Ft. \tag{8}
\]

Figure 1 depicts first-best capacity values, with the parameter values\(^7\) provided in Table 1. Both the optimal frequency and vehicle size are less than proportional to ridership. The demand elasticities of frequency and vehicle size are in the range of \(\varepsilon_F \in (0.48, 0.41)\) and \(\varepsilon_S \in (0.57, 0.65)\), respectively, as demand grows from zero to ten thousand passengers per hour. In other words, contrasting Mohring’s square root principle, vehicle size increases faster than the square root of demand, to exploit vehicle size economies.

As opposed to Jara-Díaz and Gschwender (2003), the elasticities of the optimal \(F\) and \(S\) with respect to demand add up to more than one. This implies that the optimal occupancy rate does depend on demand. As the presence of increasing returns to vehicle size suggests, high demand allows the operator to reduce the average cost of capacity provision and ease crowding under first-best conditions. This is a robust result that applies for any reasonable parameter values as long as \(\delta < 1\).

The only reason why the optimal frequency deviates from the original Mohring result is the presence of vehicle size economies. By setting \(\delta = 1\) and taking first order conditions of equation (8) with respect to \(F\) and \(S\), we get the following expressions for the optimal capacity:

\[
F = \sqrt{\frac{a}{tv}} Q \quad \text{and} \quad S = \sqrt{\frac{v \varphi \beta t}{w a}} Q. \tag{9}
\]

Now both capacity variables are proportional to the square root of demand. We can easily interpret the resulting optima. Frequency increases with the value of waiting time and decreases with the frequency related operational cost component \((v)\). Longer travel time also reduces the optimal frequency, because the share of waiting time cost falls relative to the cost of in-vehicle travel time and crowding.

\(^7\)The crowding cost parameter \((\varphi)\) is set to 0.15 as a rough approximation of the crowding cost function estimated by Hörcher et al. (2017). Travel time is now \(t = 0.25\), i.e. 15 minutes, in order to represent a standard urban public transport scenario. Of course, these values may differ significantly between public transport operators, so the primal goal of this simulation is to illustrate the mechanics of the model. The Appendix of this paper investigates model sensitivity with respect to input parameters.
Figure 1: First-best frequency, vehicle size and occupancy rate under economies of vehicle size in operational costs.

By contrast, the optimal vehicle size is inversely proportional to these values. In addition, train length increases with in-vehicle comfort related parameters and decreases with $w$ of the operational cost function. With no vehicle size economies ($\delta = 1$), the optimal occupancy rate thus becomes

$$\phi = \frac{Q}{FS} = \sqrt{\frac{w}{\varphi \beta}},$$

which is independent of demand. If vehicle size provision is expensive, then the socially optimal crowding is also higher, while high user cost parameters for in-vehicle comfort imply

---

It may be surprising that the optimal crowding level does not depend on travel time, i.e. long-distance services should offer the same level of comfort as short-haul vehicles. Note, however, that we assumed in (2) that the crowding multiplier is also independent of travel time. In other words, we neglected the fact that (standing) crowding may become more exhausting as people get tired on a long trip. Interestingly, the empirical literature of travel demand modelling has not come up yet with a convincing evidence about the exact functional relationship between $\varphi$ and $t$ (Wardman and Whelan, 2011).
lower occupancy rate in first-best optimum\(^8\). The distortion caused by the presence of vehicle size economies can be clearly observed in Figure 1. For the sake of comparison, the optimal occupancy rate at \( \delta = 1 \) is \( \phi = 1.83 \) passengers per square metre with all other parameters kept constant, which is significantly higher than the simulated results, even at the lowest level of demand.

The last panel of Figure 1 compares the average cost of waiting time \((a/F)\) to the user cost of crowding, \(\beta t \varphi Q (FS)^{-1}\). It shows that the user cost of waiting time becomes smaller in magnitude than the discomfort caused by crowding as soon as demand reaches around 5000 passengers per hour and the optimal headway drops below around 10 minutes. This simple simulation highlights the importance of crowding disutilities in supply-side optimisation of mass public transport, which has been neglected in many early models focusing on waiting time only.

2.3 The user benefits of capacity adjustment

Let us take a closer look at the distinction between waiting time and crowding costs. The original public transport model of Mohring (1972) assumes that the operator intends to minimise the sum of the total user cost of waiting time and frequency-dependent linear operational costs. The first-order condition with respect to frequency set equal to zero leads to the optimal frequency rule often called as the square root principle in which frequency is proportional to the square root of demand. Assuming that the operator is able to adjust capacity to its first-best optimum in the short run\(^9\), a marginal trip has a positive impact on frequency. This implies a marginal increase in operational costs, and a marginal reduction in waiting time for fellow passengers\(^10\). In fact, in the simple specification outlined above, the benefits enjoyed by fellow users counterbalance the operational costs of capacity adjustment, and the social cost of an incremental trip, net of personal costs, becomes zero (Small and Verhoef, 2007, Section 3.2.4). In other words, the optimal fare under marginal cost pricing is also zero. Why is this the case, and why don’t we get the same results in the presence of crowding externalities?

\(^9\)Is public transport capacity really adjustable in the short run? Throughout this paper we assume it is, just as Mohring assumed in his famous model. This is certainly a valid assumption in the planning period of a new service. There are additional reasons as well why the capacity of an existing service is more flexible than road capacity, for instance. First, ceasing operations is always a feasible option that leads to an instant reduction in operational costs, as opposed to the sunk cost of road investment. Second, for many types of public transport vehicles, there are well functioning primary and secondary markets where capacity can be purchased or sold relatively quickly.

\(^10\)Terminology: In this paper we call the impact of capacity adjustment on fellow passengers as an ‘indirect externality’. The sum of marginal personal and external user costs is referred to as the incremental ‘net user cost’. These terms can be used for waiting time and crowding costs identically. Finally, when the user cost function includes both waiting time and crowding costs, we call the sum of all direct and indirect externalities as ‘net external user cost’.
Let us define the following generalised user and operational cost functions for an unspecified capacity variable $K$, where $-\gamma_1$ and $\gamma_2$ are user and operational cost elasticities with respect to $K$.

$$\begin{align*}
C_u &= \alpha_1 Q^\kappa K^{-\gamma_1} \cdot Q; \\
C_o &= \alpha_2 K^{\gamma_2}.
\end{align*}$$

(11)

In this specification the optimal capacity, derived from first order condition $\partial(C_u+C_o)/\partial K = 0$, becomes

$$K^* = \left(\frac{\gamma_1 \alpha_1}{\gamma_2 \alpha_2} Q^{\kappa+1}\right)^{\frac{1}{\gamma_1+\gamma_2}}.$$ 

(12)

After plugging the optimal $K^*$ back into (11), we get

$$\begin{align*}
C_u &= \alpha_1 Q^{\kappa+1} \left(\frac{\gamma_1 \alpha_1}{\gamma_2 \alpha_2} Q^{\kappa+1}\right)^{-\gamma_1} \\
C_o &= \alpha_2 \left(\frac{\gamma_1 \alpha_1}{\gamma_2 \alpha_2} Q^{\kappa+1}\right)^{-\gamma_2}.
\end{align*}$$

(13)

In Mohring’s original frequency optimisation model both $C_u$ and $C_o$ are linear in capacity, so that $\gamma_1 = \gamma_2 = 1$, and $\kappa = 0$ because the cost of waiting time is independent of the number of users. In this case, the optimal capacity in equation (12) is indeed proportional to the square root of demand. Note that in any specification featuring $\gamma_1 = \gamma_2$, we find that $C_u = C_o$, i.e. total user cost equals total operational cost at all demand levels. This implies that the marginal trip with endogenous capacity has the same contribution to user and operational costs, so that

$$\frac{dC_o}{dQ} = \frac{dC_w}{dQ}.$$ 

(14)

Figure 2 relates the user benefits and operational costs of capacity adjustment to the personal waiting time cost born by the marginal user. One can validate visually in Figure 2 that condition (14), together with the optimal frequency setting rule $\frac{\partial C_o}{\partial F} = -\frac{\partial C_w}{\partial F}$, lead to simple conclusions. The external waiting time benefit, the operational cost of frequency provision, and the net impact on total (social) waiting time cost have to be equal in magnitude on the margin. Moreover, these quantities are all equal to half of the personal waiting time cost in absolute value. The latter identity will have an important role later on in this paper. Eventually, the marginal personal cost equals to the net marginal cost for society as a whole, and therefore the socially optimal fare is zero.
Figure 2: The geometric relationship between the marginal operational and marginal waiting time components in Mohring’s model. From \( \frac{\partial C_o}{\partial F} = -\frac{\partial C_w}{\partial F} \) and \( \frac{dC_o}{dQ} = \frac{dC_w}{dQ} \), it directly comes that the net waiting time effect of the marginal trip is half of the average waiting time cost.

Figure 3: The geometric relationship between the marginal operational and marginal crowding cost components, with endogenous vehicle size and exogenous frequency. From \( \frac{\partial C_o}{\partial S} = -\frac{\partial C_c}{\partial S} \) and \( \frac{dC_o}{dQ} = \frac{dC_c}{dQ} \), it directly comes that capacity adjustment fully internalises the crowding externality.

The generalised setting of (11)–(13) can be used to model crowding costs and vehicle size optimisation as well, this time with exogenous frequency. Assuming linear crowding cost and operational cost functions we set \( \kappa = 1 \), while the capacity elasticities are again equal: \( \gamma_1 = \gamma_2 = 1 \). The user cost function now represents crowding disutility, therefore we denote it with \( C_c \), but this specification is identical to the textbook case of static road congestion where the user cost function is homogeneous of degree zero. Figure 3 depicts all social costs induced by the marginal trip.

Without capacity adjustment, the marginal cost generated by an incremental trip can be split into the traveller’s own personal crowding cost, \( C_c/Q = c(Q) \), and the externality
imposed on fellow passengers, $Q \cdot c'(Q)$. Then, the operator is again able to internalise some of the newly generated user cost with capacity expansion, at the expense of operational costs. The optimal rate of vehicle size adjustment prescribes that its marginal benefit equals to its marginal cost, so that $\frac{\partial C_o}{\partial S} = -\frac{\partial C_c}{\partial S}$. As $\gamma_1 = \gamma_2$, equation (14) holds again. From Figure 3 it is clear that these two conditions cannot be met unless the crowding externality is fully internalised by the operator, so that the user benefit of capacity adjustment must neutralise the direct crowding externality entirely. The difference between the marginal social and personal costs (i.e. the optimal fare) remains equivalent in magnitude to the direct crowding externality, but this cost will actually appear in the form of an incremental operational expense.

Note that in case the crowding cost function is not homogeneous of degree zero, the average user cost may be greater or lower than the direct crowding externality: $c(Q) \neq Q \cdot c'(Q)$. Greater personal cost, for example, would imply that not just the crowding externality can be internalised, but crowding will actually ease as a result of capacity adjustment on the margin. As a consequence, the optimal fare (social cost minus personal cost) becomes lower than the marginal operational cost, and the marginal trip will have to be subsidised. In summary, with crowding costs and endogenous vehicle size, the marginal personal cost of travelling is just a fraction of its marginal social cost, and therefore the rest has to be internalised with pricing. In Section 3.1.3 we discuss the case of simultaneous adjustment of frequency and vehicle size, with waiting time as well as crowding on the user cost side.

2.4 Cost Recovery Theorem for public transport

The optimal degree of subsidisation, in other words cost recovery, is one the key policy questions in public transport. We briefly discuss the adaptation of Cost Recovery Theorem (CRT) to crowding and waiting time costs. In the road literature Mohring and Harwitz (1962) showed that optimal static congestion charging leads to cost recovery ratio

$$\eta = \varepsilon + h \cdot \frac{Q c_u(Q,K)}{C_o(K)},$$

(15)

if (i) the user cost function $c_u(Q,K)$ is homogeneous of degree $h$, (ii) the elasticity of the cost of capacity provision is $\varepsilon$, and (iii) capacity is indivisible. In the specific case when the congestion delay only depends on the ratio $Q/K$ (so that $h = 0$), and investment costs feature constant returns to road capacity ($\varepsilon = 1$), optimal pricing leads to full cost recovery. Small and Verhoef (2007, Section 5.1.1) show that the cost recovery result holds for multi-period settings as well, as long as $h = 0$.

Let us now consider frequency optimisation in public transport with waiting time and crowding disutilities. With a linear additive specification the average user cost function becomes $c_u(Q,F) = k_1 \cdot F^{-1} + k_2 \cdot QF^{-1}$. Imagine first that crowding costs are negligible
compared to waiting time costs \((k_1 >> k_2)\). What we get is the standard Mohring result. The user cost function is homogeneous of degree \(-1\), as \(c_u(\lambda Q, \lambda F) = \lambda^{-1} c_u(Q, F)\). Assuming that operational costs feature constant returns to scale \((\varepsilon = 1)\), equation (15) correctly suggests that all operational costs have to be covered with public subsidies. By contrast, if \(k_2 >> k_1\), so that the inconvenience of waiting is negligible compared to crowding, we get back to the original model of transport supply with constant returns to scale in user costs and full cost recovery. It is clear, however, that the baseline CRT assumptions may not hold in public transport. If both user cost components are non-negligible, as it is normally the case in reality, \(c_u(Q, F)\) is no longer homogeneous, and therefore equation (15) of the CRT cannot be applied any more to determine the cost recovery ratio.

3 Demand imbalances and second-best performance

The rest of this paper discovers the impact of demand imbalances in a multi-period supply regime on the economic efficiency and financial performance of public transport provision. We show that beside the structure of user costs, the optimal subsidy depends on the pattern of demand as well. This section is split into two parts: Section 3.1 restricts the analysis to inelastic demand, where capacity setting boils down to a cost minimisation problem. The main benefit of the inelastic demand assumption is that some of the numerical simulation results can be reinforced with analytical derivations. Then Section 3.2 provides additional insights by considering structural differences between the underlying demand curves when unbalanced ridership levels can be observed in equilibrium. In both sections, we begin the analysis with the derivation of optimal supply variables in function of the degree of demand asymmetry, and then focus on economic and financial performance.

3.1 Fixed demand

3.1.1 Second-best capacity

Let us consider the simplest form of demand imbalances: the back-haul problem. Service provision from A to B implies identical capacity supply from B to A as well, while demand can be unbalanced between the two directions. We define \(Q_1\) and \(Q_2\) as constant ridership in the busy and calm directions, respectively, such that \(Q = Q_1 + Q_2\). We keep the cost specifications of equation (2), and express social cost as

\[
\min_{F, S} TC(F, S, Q_1, Q_2) = aF^{-1}(Q_1 + Q_2) + \beta t(Q_1 + Q_2) + \beta t \varphi \frac{Q_1^2 + Q_2^2}{FS} + 2tF(v + wS^d).
\]

Note that the two demand levels, \(Q_1\) and \(Q_2\) enter the crowding cost function as quadratic terms, while the user cost of waiting time and the uncrowded travel time are linear in aggre-
gate demand. This suggests that the more imbalanced the two markets are, i.e. the more \( Q_1 \) and \( Q_2 \) deviate from \( Q/2 \), the more important the user cost of crowding will become.

In order to express the magnitude of demand imbalance in a single variable, we define \( \omega = Q_1/Q \) as the share of peak demand in the aggregate passenger flow. This may vary between 0.5 and 1, i.e. between a symmetric setting and a completely unidirectional demand pattern at the two extremes. We fix total demand at \( Q = 5000 \) passengers per hour, which allows us to investigate the pure implications of \( \omega \) without any aggregate scale effects. Figure 4 presents simulation results for the optimal second-best frequency and vehicle size as a function of the magnitude of demand imbalances (\( \omega \)), with the vehicle size elasticity parameter set to \( \delta = 0.8 \).

**Figure 4:** Second-best optimum of vehicle size and frequency provision as a function of the magnitude of demand imbalance, i.e. the share of peak demand (\( \omega \)). Aggregate demand is fixed at \( Q = 5000 \, \text{pass/h} \).

The immediate observation is that \( \omega \) has a negative impact on frequency and a positive effect on vehicle size. Intuitively, as more people are concentrated in one of the two directions,
the more important it becomes to supply sufficiently spacious vehicles, even at the expense of lower frequency and higher expected waiting time costs. The analytical derivation of the optimal capacity from equation (16) provides more insights. This, again, is only feasible with $\delta = 1$, in which case we get

$$F = \sqrt{\frac{a}{2tv} Q} \quad \text{and} \quad S = \sqrt{\frac{v \varphi \beta t}{w} \rho Q},$$

where $\rho = \omega^2 + (1 - \omega)^2$. (17)

It turns out that the second-best frequency is independent of the magnitude of demand imbalances when vehicle size economies do not exist. In fact, the second-best frequency is identical to the first-best expression in equation (9) applied for the average ridership, which is $(Q_1 + Q_2)/2$ in our case. Vehicle size, on the other hand, does depend on the magnitude of demand imbalances through coefficient $\sqrt{\rho}$, which is an increasing function of $\omega$. This reliance on vehicle size provision becomes more significant when density economies exist. Due to the fixed cost of frequency provision, the marginal operational cost of frequency is higher than the marginal cost of vehicle size adjustment, and therefore it is more efficient to address crowding-related user costs by vehicle size adjustment only. This provides a second explanation for the shift from frequency to vehicle size provision.

As total capacity adds up to

$$FS = Q \sqrt{\frac{\varphi \beta \rho}{w} \frac{1}{2}},$$

it can be concluded that total capacity is proportional to aggregate demand ($Q$) and increases with the asymmetry between the two markets. This suggests that the average operational and environmental cost of a trip is higher in the presence of heavy demand fluctuations.

Let us return to Figure 4 and the case of vehicle size economies, and investigate the resulting second-best occupancy on the two markets. Unsurprisingly, crowding decreases in the back haul, simply because passengers disappear as we move towards strong demand imbalances. Even though the total second-best capacity increases with $\omega$, crowding on the busy direction still grows, with a decreasing rate. The average occupancy rate on the two services slightly reduces with $\omega$ on a per train basis. However, as Rietveld (2002) pointed out correctly, this operational average is not what passengers experience in reality. The average crowding density per passenger, measured by

$$\bar{\phi}_p = \frac{Q_1 \phi_1 + Q_2 \phi_2}{Q_1 + Q_2}$$

increases with the difference in demand between the two markets. Indeed, when the off-peak train is empty, i.e. $\omega = 1$, the average occupancy rate per train is half of what passengers experience. This casts doubts on whether the average occupancy rate per vehicle, a performance indicator often used by public transport operators, is a valid measure of service quality.
3.1.2 Marginal social cost components

Marginal social costs have important implications on optimal pricing and subsidisation. It is therefore worth exploring the differences in the marginal cost pattern of peak and off-peak markets. The difference between marginal social cost on market $i$ and the average user cost on the same market gives the net social burden of the marginal trip imposed on fellow users and the operator:

$$MEC_i = \frac{\partial TC}{\partial Q_i} - c_u(Q_i) = a(Q_1 + Q_2)\frac{\partial F^{-1}}{\partial Q_i} + \frac{Q_i}{FS}_\text{DCE} + \frac{\varphi(Q_1^2 + Q_2^2)}{\partial Q_i}_\text{ICE} \beta t + \frac{Q_i}{FS} \beta t \frac{(FS)^{-1}}{\partial Q_i} b_t$$

$$+ (v + wS^\delta)\frac{\partial F}{\partial Q_i} + Fw\delta S^\delta - 1 \frac{\partial S}{\partial Q_i}.$$

Equation (20)

First, the marginal consumer imposes a direct crowding externality (DCE) on fellow users, simply due to the fact that more passengers have to share the available in-vehicle floor area. Second, the operator adjusts capacity supply in order to keep it at its optimal level\(^{11}\). As we saw in the previous section, both the optimal frequency and vehicle size react to an increase in demand. As a result, marginal operational costs (MOC) will be generated, and the capacity expansion/reduction will have an impact on fellow passengers’ user costs as well. In particular, the marginal change in headways, $\partial F^{-1}/\partial Q_i$, affects waiting times for customers in all markets, and the deviation of total in-vehicle capacity, $\partial (FS)^{-1}/\partial Q_i$, has an influence on crowding experience as well. The latter components are marked as indirect waiting time (IWE) and crowding (ICE) externalities in equation (20).

If capacity variables are set optimally, the envelope theorem assures that the sum of the operational cost and indirect externality components is zero on the margin. That is, the marginal cost of capacity adjustment is equal to its marginal benefit, as equations (5) and (6) suggest. Nevertheless, the magnitude of these components is still relevant, from the viewpoint of subsidisation, for instance.

Figure 5 shows the sign and magnitude of frequency and vehicle size adjustment on the margin. The incremental trip increases the aggregate demand level ($Q$) and induces a shift in the relative size of peak and off-peak markets ($\omega$ changes). In the previous section we found that both capacity variables react positively to an increase in $Q$, while $\partial F/\partial \omega < 0$ and $\partial S/\partial \omega > 0$. As a result, the marginal off-peak trip provokes stronger frequency adjustment than the marginal peak trip. The opposite applies for the marginal vehicle size. Moreover, at high demand imbalances, the marginal off-peak trip may even reduce the optimal second-best vehicle size through the increase in $\omega$. At extremely heavy demand imbalances, total capacity

\(^{11}\)For the sake of simplicity we did not indicate that capacity is optimised in this cost function, so that $F = F(Q, S)$ and $S = S(Q, F)$. 

18
may also decrease (at the expense of vehicle size solely, because the marginal frequency can never be negative).

Figure 5: Change in the optimal frequency, vehicle size and overall capacity after a marginal increase in peak (solid lines) and off-peak (dashed lines) demand, as a function of $\omega$.

Figure 6: Marginal social cost components in the back-haul setting. Dashed lines represent the off-peak market.

After plugging these marginal capacity values back to the cost expressions of equation (20), we get the marginal cost pattern depicted in Figure 6. The first panel compares the direct and indirect externalities of crowding, the second plots the resulting net crowding effect and the indirect benefit (negative cost) of waiting time reduction, while the third panel adds up all user externalities and shows their magnitude relative to the marginal operational cost induced by capacity expansion. Panel 6a clearly indicates that the direct and indirect crowding externalities almost cancel each other out. Indeed, when $\delta = 1$, external crowding costs and benefits completely neutralise each other, just like in the simple example of Section 2.3. The slight difference in marginal crowding externalities have to be attributed to vehicle size economies. Peak trips on the margin induce more indirect crowding benefits,
as they allow the operator to exploit vehicle size economies more effectively. Waiting time externalities show the opposite pattern: off-peak trips generate more benefits because they induce stronger positive frequency adjustment through the increase in $\omega$.

Unsurprisingly, peak trips induce much higher marginal operational costs than off-peak users. The difference equals to the difference in the direct crowding externality they cause. The net social cost, i.e. the optimal fare in case of marginal social cost (MSC) pricing, boils down to this direct crowding externality, which is in fact born by society in the form of incremental operational costs. This suggests that, in case of MSC pricing, peak and off-peak trips should receive the same public subsidy on the margin. In Section 3.2.2 we show however that total subsidies may be fundamentally different for the two markets.

### 3.1.3 Uniform marginal subsidy for peak and off-peak trips

The net user externality of peak and off-peak trips is identical in Figure 6, no matter how strong the imbalance between the two markets is. That is, peak and off-peak trips should receive the same marginal subsidy under optimal pricing. This, of course, is not a coincidence. Numerically, this common net marginal user externality (MUE) equals in magnitude to half of the average waiting time cost, i.e. $\text{MUE} = -0.5 a F^{-1}$. To better understand the mechanics behind this result, recall our analysis of the basic Mohring model in Section 2.3. Recall that Figure 2 shows that the marginal waiting time benefit generated by optimal frequency adjustment equals to half of the personal waiting time cost.

Let us plot the current relationship between marginal user and operational costs in a similar graph in Figure 7. The main driving forces of the mechanism are still (i) the optimality condition for capacity setting prescribing $\frac{\partial C_o}{\partial (FS)} = -\frac{\partial C_u}{\partial (FS)}$, and (ii) $\frac{\partial C_o}{\partial Q_i} = \frac{\partial C_u}{\partial Q_i}$ resulting from equal user and operational cost elasticities with respect to capacity (see Secion 2.3). That is, the two pairs of grey and dashed vectors in the figure need to have the same length to reach the optimum.

Note that the change in total user cost due to capacity adjustment, $\frac{\partial C_u}{\partial (FS)}$, now contains the benefits enjoyed by passengers outside of market $i$ as well, while the marginal trip occurs in market $i$. The graph does not detail how these benefits are distributed among jointly served markets, but in fact it does not matter for the purpose of this analysis. It is more important that if the crowding cost function is homogeneous of degree zero, so that if the average user cost of crowding equals to the direct crowding externality, then the net user externality, calculated as $\frac{dC_u}{dQ_i} - c_i(Q_i)$, becomes half of the personal waiting time cost. This applies for all markets of the network. As the personal cost of waiting is the same in both peak and off-peak markets while the rest of the graph is symmetrical to the horizontal axis at 0.5, the net external user cost is also identical in both directions.
3.2 Elastic demand

What we observe in reality as differences in actual passenger numbers is the just the equilibrium outcome of the transport market, as defined by equation (3). What is in fact exogenous to the public transport operator is not the actual ridership on jointly served markets, but the differences between the underlying demand functions. In this section we consider varying inverse demand curves in the two markets of the back-haul setting, such that \( d_1(q) \neq d_2(q) \).

With elastic demand, it is not straightforward how we define the notion of demand imbalances when comparing two functions, as the asymmetry in equilibrium ridership is now endogenous. In order to keep the analysis tractable, we restrain our attention to linear inverse demand curves and distinguish three idealised cases:

- **MS**: unbalanced market size with fixed maximum willingness to pay;
- **WTP**: unbalanced maximum willingness to pay with fixed market size\(^{12}\);
- **MIX**: varying market size and maximum willingness to pay with constant generalised price sensitivity of demand: \( d_1'(q) = d_2'(q) \).

Up until this point, the analysis follows the approach of Bichsel (2001) who investigated the consequences of demand imbalances with uniform road tolls. Furthermore, in order to

---

\(^{12}\)The spatial and temporal variability of maximum willingness to pay may stem from various sources in a transport context. It can be related to income inequality, variety in preferences, and heterogeneity in the fixed cost of accessing the transport system, among others.
make these cases comparable by controlling for scale effects, we keep the sum of aggregate consumer benefits on the two markets constant at $A_0M_0$. That is, total willingness to pay for the costless services remains constant, and $\theta$ is defined as the share of peak aggregate consumer benefit:

$$\theta = \frac{B_1}{B_1 + B_2}, \quad \text{where} \quad B_i = \int_0^\infty d_i(q) dq \quad \text{and} \quad B_1 + B_2 = A_0M_0.$$  \hfill (21)

Said differently, the measure of demand imbalances in the elastic demand context is now the difference in total willingness to pay in the two markets, which may or may not result in differences in ridership$^{13}$ ($\theta \not\equiv \omega$). Figure 8 provides a graphical overview of the three cases, and Table 2 details how peak and off-peak inverse demand curves are generated in accordance with equation (21). This setting enables the investigation of the impact of unbalanced demand on economic efficiency as well.

![Figure 8: Idealised models of market size and willingness to pay imbalances with elastic demand, in the back-haul problem. Aggregate consumer benefit on the two markets (i.e. the area below the inverse demand curves) is kept constant.](image)

<table>
<thead>
<tr>
<th>Case</th>
<th>Market</th>
<th>$A$ (max. w.t.p)</th>
<th>$M$ (market size)</th>
<th>Slope of demand curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>Peak</td>
<td>$A_0$</td>
<td>$2\theta \cdot M_0$</td>
<td>$-[2\theta]^{-1} \cdot A_0/M_0$</td>
</tr>
<tr>
<td></td>
<td>Off-peak</td>
<td>$A_0$</td>
<td>$2(1-\theta) \cdot M_0$</td>
<td>$-[2(1-\theta)]^{-1} \cdot A_0/M_0$</td>
</tr>
<tr>
<td>WTP</td>
<td>Peak</td>
<td>$2\theta \cdot A_0$</td>
<td>$M_0$</td>
<td>$-2\theta \cdot A_0/M_0$</td>
</tr>
<tr>
<td></td>
<td>Off-peak</td>
<td>$2(1-\theta) \cdot A_0$</td>
<td>$M_0$</td>
<td>$-2(1-\theta) \cdot A_0/M_0$</td>
</tr>
<tr>
<td>MIX</td>
<td>Peak</td>
<td>$\sqrt{2\theta} \cdot A_0$</td>
<td>$\sqrt{2\theta} \cdot M_0$</td>
<td>$-A_0/M_0$</td>
</tr>
<tr>
<td></td>
<td>Off-peak</td>
<td>$\sqrt{2(1-\theta)} \cdot A_0$</td>
<td>$\sqrt{2(1-\theta)} \cdot M_0$</td>
<td>$-A_0/M_0$</td>
</tr>
</tbody>
</table>

These idealised demand specifications allow us to solve the welfare maximisation problem.

---

$^{13}$That is, unlike in Section 3.1, total demand may not remain constant in equilibrium as we increase the magnitude of asymmetry.
of equation (4) for four decision variables: peak fare \((p_1)\), off-peak fare \((p_2)\), frequency \((F)\) and vehicle size \((S)\). As the analytical derivation of optimal supply is unfeasible in this section, we use five prior observations to ease the interpretation of simulation results.

O1 Section 3.1 revealed that the concentration of demand in one market induces a shift from frequency to vehicle size provision. This effect should persist with elastic demand as well.

O2 Bichsel (2001) analysed the financial performance of MS and MIX scenarios for the case of uniform second-best pricing with fixed capacity. In the MS case it is the main haul that becomes more elastic, as opposed to the other regimes where off-peak demand becomes more sensitive. Bichsel (2001) shows that the Ramsey mechanism leads to better-than-first-best financial result in the MS case, as the second-best toll moves towards the first-best toll of the sensitive market. By contrast, in the opposite case with high off-peak demand sensitivity, the profit outcome is worst than first-best. Although in our study prices are allowed to differ between markets, we expect that the same mechanism could work with the two second-best capacity variables as well.

O3 The simple case of a horizontal (constant) marginal cost curve provides additional prior insights. As long as the generalised price is less than the maximum willingness to pay \((A_0)\), demand in the off-peak market can never drop to zero in the MS setting, even amid high demand imbalances. In the WTP and MIX cases, on the other hand, it is possible that off-peak services remain unused if the highest willingness to pay falls below the cost of travelling in this market. Moreover, it is easy to recognise geometrically as well as analytically that aggregate equilibrium demand in the two markets is independent of \(\theta\) under MS inequality, but it strictly decreases with \(\theta\) in the WTP model.

O4 In the same setting with constant marginal cost, aggregate consumer surplus also remains independent of the degree of imbalance in the MS scenario, while it is a decreasing function of \(\theta\) in WTP. Thus, one may expect a priori lower consumer surplus in the latter scenario under endogenous capacity as well.

O5 However, there is a major difference in the way how \(\theta\) increases total willingness to pay in the peak market: in the MS case it increases the number of potential (price sensitive) users by adding more willingness to pay to the right hand side of the inverse demand distribution, while in WTP it adds more willingness to pay to existing users, thus making them less sensitive to changes in travel costs.

3.2.1 Optimal supply

In all subsequent simulation exercises we found that the MIX scenario with constant generalised price sensitivity provides intermediate results between the extreme values of fixed willingness to pay and market size. Therefore we focus on the latter cases in the interpretation of our findings.
Optimal capacity is indeed a function of the magnitude of demand imbalances, as Figure 9 depicts. Moreover, alternative demand specifications lead to unequal optima, suggesting that the source of demand imbalances is indeed a relevant issue to analyse. In line with Section 3.1 and prior observation O1, as $\theta$ grows, the operator will reduce frequency and substitute capacity with larger vehicles. Frequency reduction is much stronger in the WTP case, but the corresponding vehicle size compensation is limited. This can be explained as the combined effect of O1 and O3. Figure 10.c justifies O3 in the sense that aggregate demand is almost constant in the MS case, while it strongly decreases in the WTP scenario. As a consequence, beside the shift from frequency to vehicle size provision, both capacity variables fall in the WTP model due to the reduction in aggregate equilibrium demand. This leads to strictly decreasing frequency and ambiguous changes in vehicle size as $\theta$ increases.

Additional optimal pricing and ridership results are provided in Figure 10. As one would expect, at very high $\theta$ the operator does not does not face demand in the off-peak market at all\textsuperscript{14}, while in the MS case $Q_2 > 0, \forall \theta < 1$. The mixed setting plays an intermediary role again in the sense that the $\theta$ value at which $Q_2$ drops to zero is larger than in the WTP model but lower than unity. Not only the sum of passenger numbers, but also the differences between peak and off-peak demand and crowding levels are the highest when market size varies. According to the pricing formula we derived in equation (7), optimal fares in Figure 10.a are proportional to the density of crowding. The Ramsey mechanism of expectation O2 suggests that the second-best capacity moves towards what the first-best would be in the

\textsuperscript{14}In all figures of this section we indicate the critical level of $\theta$ above which the back haul is not served any more with vertical dashed lines. The off-peak market not being served means here that equilibrium demand drops to zero, i.e. potential users' willingness to pay remains below the uncrowded generalised user cost, so that back-haul services are running empty.
more elastic market: in the MS case this implies higher capacity to accommodate sensitive peak demand and thus low crowding in the back-haul, and a narrower gap in crowding levels amid varying WTP. Simulation results justify this expectation. This second-best mechanism assures that aggregate demand is always relatively higher in the more sensitive market.

Figure 10: Welfare maximising fares and the resulting equilibrium demand levels in peak (solid lines) and off-peak (dashed) markets.

Figure 11: Average occupancy rate on a per passenger and per train basis.

The impact of demand imbalances on the crowding experience for the average passenger is strictly negative, as Figure 11 depicts. Compared to the MS scenario, average crowding only slightly increases at low levels of \( \theta \) amid willingness to pay imbalances. Later on, however, this dramatically increases until the off-peak market completely disappears. This is a joint implication of phenomena O2 and O3: although the peak market is always more sensitive in the MS scenario, \( \theta \) leads to a reduction in aggregate demand in the WTP regime, thus allowing for better crowding conditions in optimum. Finally, in line with the inelastic demand case, average train occupancy rate statistics show a misleading, opposite pattern, with the lowest crowding levels for the MS case.
3.2.2 Efficiency and subsidisation

From a policy perspective, relevant conclusions can be drawn from the simulation results by looking at the maximum social surplus, specified in equation (1), that can be achieved depending on the strength of $\theta$. The corresponding financial result is the difference between revenues, $R = \sum_i p_i \cdot Q_i (c_{u,i} + p_i)$, and the operational cost, i.e. $C_o$ in equation (2).

![Diagram](image.png)

**Figure 12:** Social welfare in optimum and the resulting financial result. Negative financial results imply subsidisation.

Recall that with no user costs at all, the three models would lead to the same aggregate consumer benefit (total willingness to pay) measured by $A_0 M_0$. We found in the previous subsection that simulation results are in line with expectation O3 in terms of the equilibrium aggregate demand. By contrast, Figure 12 suggests that expectation O4 does not hold when it comes to consumer surplus and social welfare: as $\theta$ grows, social welfare decreases in the MS scenario, and increases at the WTP model. Taking a closer look at the main components of social welfare in Figure 12, one can notice that the majority of improved social welfare
in the WTP scenario comes from two sources: (i) peak consumer surplus increases heavily, which is a straightforward consequence of prior observation O5, and (ii) operational expenses decrease, which comes from the reduction in aggregate demand and the corresponding savings in optimal capacity provision. These two effects are more significant in magnitude than the accelerating reduction in off-peak consumer surplus. In other words, the more severe the difference in willingness to pay between the two directions, the more social welfare can be generated by public transport provision, *ceteris paribus*\textsuperscript{15}.

Finally, the financial result in Figure 12.b is also counter-intuitive for the first sight. According to Bichsel (2001), the MS scenario achieves better-than-first-best cost recovery result in a multi-period road supply model. What we find for the case of public transport is the opposite: under market size imbalances the service should receive more public subsidy to maximise efficiency than under willingness to pay imbalances. This major difference can be explained by the presence of positive waiting time externalities that prescribe subsidisation by default. Section 3.1 found that the marginal passenger’s subsidy should remain within a narrow range in all markets served by the same capacity. The pattern of optimal subsidies, in other words the mirror image of Figure 12.b, resembles the pattern of aggregate demand in the two scenarios: as aggregate demand decreases in the WTP scenario, it is a natural consequence that the total subsidy should also decrease. Contrasting the WTP case, the financial result slightly further deteriorates under market size inequality. A sensitivity analysis discussed in the Appendix revealed that the moderately degrading financial result can be attributed to the subsidisation of vehicle size economies. With $\delta$ set close to 1, the optimal subsidy may stay constant or even slightly decrease with $\theta$, just as aggregate peak demand slightly decreases in Figure 10.

In Table 3 we summarise the main results of the analysis presented in this section. With the exception of vehicle size, all supply variables as well as the resulting economic measures are monotonous functions of the magnitude of demand imbalances. Therefore the slope of these functions give a reliable indication of how $\theta$ affects them.

Many findings of the elastic demand analysis in Section 3.2 are based on numerical simulations only. In order to ensure the robustness of the results with respect to input parameters, we perform a sensitivity analysis using quantitative as well as descriptive methods in the Appendix of this paper. The sensitivity analysis concludes that reasonable variations in input parameters do not affect the qualitative results summarised in Table 3.

\textsuperscript{15}The reader should note that this statement holds under constant aggregate consumer benefit only. Our results do not imply that policies that generate heterogeneity in maximum willingness to pay can improve social welfare *per se*. It is always worth filling up empty spaces in the back haul if peak demand can be kept constant, thus increasing the overall scale of total willingness to pay.
Table 3: Summary of results.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>MS</th>
<th>Relation</th>
<th>WTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>Peak fare</td>
<td>+</td>
<td>$&gt;$</td>
<td>+</td>
</tr>
<tr>
<td>$p_2$</td>
<td>Off-peak fare</td>
<td>$-$</td>
<td>$&lt;$</td>
<td>$-$</td>
</tr>
<tr>
<td>$F$</td>
<td>Frequency</td>
<td>$-$</td>
<td>$&gt;$</td>
<td>$-$</td>
</tr>
<tr>
<td>$S$</td>
<td>Vehicle size</td>
<td>+</td>
<td>$&gt;$</td>
<td>ambiguous</td>
</tr>
<tr>
<td>$Q_1$</td>
<td>Peak demand</td>
<td>+</td>
<td>$&gt;$</td>
<td>+</td>
</tr>
<tr>
<td>$Q_2$</td>
<td>Off-peak demand</td>
<td>$-$</td>
<td>$&lt;$</td>
<td>$-$</td>
</tr>
<tr>
<td>$Q$</td>
<td>Total demand</td>
<td>$-$</td>
<td>$&gt;$</td>
<td>$-$</td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>Peak crowding</td>
<td>+</td>
<td>$&gt;$</td>
<td>+</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>Off-peak crowding</td>
<td>$-$</td>
<td>$&lt;$</td>
<td>$-$</td>
</tr>
<tr>
<td></td>
<td>Average crowding per passenger</td>
<td>+</td>
<td>$&gt;$</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Average crowding per train</td>
<td>$-$</td>
<td>$&lt;$</td>
<td>$-$</td>
</tr>
<tr>
<td></td>
<td>Peak consumer surplus</td>
<td>+</td>
<td>$&lt;$</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Off-peak consumer surplus</td>
<td>$-$</td>
<td>$&gt;$</td>
<td>$-$</td>
</tr>
<tr>
<td>$B - C_o$</td>
<td>Total consumer surplus</td>
<td>$-$</td>
<td>$&lt;$</td>
<td>+</td>
</tr>
<tr>
<td>$C_o$</td>
<td>Operational costs</td>
<td>+</td>
<td>$&gt;$</td>
<td>$-$</td>
</tr>
<tr>
<td>$SW$</td>
<td>Social welfare</td>
<td>$-$</td>
<td>$&lt;$</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Optimal subsidy (financial loss)</td>
<td>+</td>
<td>(*)</td>
<td>$&gt;$</td>
</tr>
</tbody>
</table>

*Signs indicate monotonous increase or reduction in function of $\theta$.

Relation signs compare the optimal values in MS and WTP models.

* At low $\delta$ the optimal subsidy may decrease with $\theta$: see Appendix.

4 Future research

The pure back-haul setting is rare in public transport as spatial and temporal demand variations may be equally important in a network. Therefore it is an objective of great relevance for future research to generalise the main findings derived from the back-haul setting. As an example, let us consider the case of $m = 4$ markets served by the same second-best capacity. Having only one peak market with three off-peak sections would increase the share of calm markets to 75% in operational costs, assuming identical travel time on all segments. This implies that accommodating peak demand becomes more expensive compared to the original back-haul case. In the MIX demand model with unchanged parameters (see Table 1), the highest level of peak crowding increases from 1.32 to 1.88 passengers per square meter. The corresponding peak fare increases from 0.99 monetary units to 1.41. In other words, not only the peak market’s share in aggregate demand ($\omega$, or in case of elastic demand $\theta$) matters, but also the peak market’s share in total network length. Theoretically the optimal second-best crowding density may reach extremely high levels, given that a peak market has very high share in total demand, and very low share in network length.
Taking spatial as well as temporal demand imbalances into consideration, several hundreds of markets may have to be served by the same capacity. In such a setting we cannot clearly distinguish peak and off-peak markets; a more inclusive measure of demand imbalances should be applied to determine how the demand pattern affects the economic performance of public transport. A suitable measure may be a Gini coefficient of demand inequality. The Lorenz curve in this case would express the cumulative share of demand (or consumer benefit) as a function of the cumulative share in network length (travel time or distance). Future research should address the straightforward question of how this index affects public transport performance. The preliminary analysis of a simple network shows that for a sample of randomly generated OD matrices with fixed output, measured by the number of passengers as well as passenger-kilometers, the Gini coefficient is a surprisingly good predictor of the optimal frequency, vehicle size, operational costs and average crowding experience. The shape of the relationships between these variables and the Gini coefficient are very similar to what we found in the back-haul setting.

Crowding generates demand interactions between overlapping origin-destination markets. Yang and Meng (2002) derive that link-based marginal external cost tolls are efficient in a road network, assuming that user costs are homogeneous of degree zero. As this assumption does not hold public transport setting, optimal crowding pricing remains an open question in such networks. Additionally, the technological constraint that several markets have to be served by constant capacity is in a way endogenous, as operators do have control over how the network of lines of various transport modes is organised. Thus, it is clear that network design is also severely affected by the pattern of spatio-temporal demand imbalances in an urban area.

Beside the extension to complex networks, this paper leaves considerable room to investigate more sophisticated pricing policies under unbalanced demand. For example, urban public transport operators usually cannot differentiate fares between directions, or even between line sections and time periods. The usual tariff rules, allowing flat, distance based or zone based fares only, impose additional constraints on the supply optimisation problem discussed in this paper. Also, the welfare maximisation objective is often exogenously constrained by the limited availability of public subsidies. Jara-Díaz and Gschwender (2009) suggest that the budget constraint leads to sub-optimal capacity provision. Future research may investigate how the magnitude of demand imbalances affect these ‘third-best’ optima when setting fares and capacities under limited financial resources.

Our analysis completely lacks any interaction with competing modes, most importantly with individual car transport. It is likely that demand imbalances in public transport coincide with similar phenomena in other modes, especially if these modes are substitutes. In the presence of unpriced consumption externalities, public transport supply has to react to demand imbalances in other modes. For the specific case of untolled peak road usage and
public transport, intuition suggests that peak fares should be lower than what we found in this paper to mitigate some of the deadweight loss of the congestion externality (Anderson, 2014, Tirachini et al., 2014).

We do not investigate in the present study the behaviour of a profit oriented monopolist in fluctuating demand, nor any market outcomes in a competitive framework. Nevertheless, the analyses presented above could be reconsidered with a profit maximising objective as well\textsuperscript{16}. Theory highlights that profit maximising operators also internalise the external cost of crowding, and therefore one may expect dynamic fares correlated with the density of crowding. This may be distorted, however, by the monopoly mark-up which depends on the price sensitivity of demand. In this case the source of demand imbalances, i.e. the slope of peak and off-peak inverse demand curves, would play an important role.

The wide range of potential topics suggests that the analysis of second-best capacity provision and crowding pricing in public transport may become an emerging field in the future.

5 Conclusions

The primal aim of this paper is to draw attention to the importance of multi-period supply optimisation in public transport service provision. We show that the usually unavoidable property that the same capacity has to serve multiple spatio-temporally differentiated markets implies second-best capacity provision, if these markets feature nonidentical demand conditions. The magnitude of the difference in demand has a number of consequences on the economic and financial performance of the service.

The contribution of this paper is twofold: we add theoretical as well as practical insights to the existing literature. On the theory side we analyse the interplay between external waiting time benefits and crowding costs with endogenous capacity. The paper discusses the application of the Cost Recovery Theorem with waiting time and crowding costs. Also, we reveal that the primal consequence of unbalanced demand is the shift from frequency to vehicle size oriented capacity provision. On the practical side, we show that crowding in peak markets may well be an optimal outcome of second-best capacity provision under demand fluctuations. That is, peak crowding is not necessarily a sign of malfunctions in supply.

After decomposing the external user and operational cost components of the marginal trip, we derive that despite the difference in operational costs, the marginal peak and off-peak trip should receive the same monetary subsidy. We prove in an extended graphical representation of Mohring’s original capacity model that the uniform marginal subsidy equals to half of the

\textsuperscript{16}The foundations of the back-haul problem in a profit maximising framework with peak ridership dependent joint costs have been laid down by Mohring (1970) and Rietveld and Roson (2002).
average waiting time cost in all markets, if the crowding cost function is homogeneous of
degree zero, so that marginal personal and external crowding costs are equal.

In an elastic demand setting we disentangle the effect of two major sources of demand im-
balances: variations in market size and maximum willingness to pay. The analysis concludes
that the source and the magnitude of the imbalance may affect not only the optimal supply
variables, but also the attainable level of social surplus and the operator’s financial result.
In particular, we find that amid pure market size imbalances the expected social welfare
decreases, and in case of considerable vehicle size economies, the optimal subsidy increases
with the strength of unbalancedness. The opposite results are reported for pure willingness
to pay deviations, low demand users being identified as the victims of demand imbalances.

Although the paper deals with the simplest network layout, the back-haul problem, this
research opens up a wide range of topics on the interrelations between transport services and
the urban economy. Demand imbalances are rooted in the heterogeneity in urban spatial
structure and the temporal sequence of activities. Our results suggest that spatial policies
may have considerable impact on the economic and financial performance of public transport.
Therefore, when benchmarking such services, one should take into consideration the urban
environment in which they are operated.

Acknowledgement

The authors thank all insightful comments at the ITEA (Santiago de Chile), ERSA (Vienna,
Austria) and NARSC (Minneapolis, U.S.) conferences in 2016. We are especially grateful
to Erik Verhoef, Fangni Zhang and three anonymous referees for the in-depth feedback they
provided.

Appendix. Sensitivity analysis

Global sensitivity analysis

The optimal supply variables in Section 3.2 are in line with the patterns we found in Sec-
tion 3.1 with inelastic demand, and those results can be verified with analytical calculations.
Social welfare and optimal subsidy functions, on the other hand, are derived from numerical
simulations only, and therefore the qualitative conclusions summarised in Table 3 may be sen-
tsitive to input parameters. In this sensitivity analysis below we thus focus on the robustness
of social welfare and financial results of Section 3.2.

Maximum (optimal) social welfare and the resulting financial outcome are expressed as
functions of seven main input parameters: $a, \beta, t, \phi, v, w$ and $\delta$. We apply two global sensi-
tivity analysis methods to identify the input parameters that the two outputs may be the
most sensitive to. Figure 13 depicts the results of our investigation with the Morris method and Figure 14 provides Sobol indices (Iooss and Lemaître, 2015, Saltelli et al., 2008).

Figure 13: Sensitivity screening of the MIX model with Morris method at $\theta = 0.75$, with $r = 50$ OAT designs and $l = 5$ discrete input levels.

Morris’ screening method is based on a series of one input at a time (OAT) random experiments which allows the measurement of the elementary effect of each input parameter on the output values (Morris, 1991). The distribution of the absolute value of the elementary effect is described by its mean ($\mu_i^*$) and standard deviation ($\sigma_i$) for input variable $i$ (Campolongo et al., 2007). More specifically, $\mu_i^*$ is a good proxy for the magnitude of the linear effect of the input, while $\sigma$ measures the strength of nonlinear and interaction effects.

From Figure 13 one can infer that $\beta$ and $t$ have the strongest influence on the attainable social welfare among the input parameters. Travel time ($t$) has a predominantly linear effect, while the uncrowded value of time $\beta$ is close to the $45^\circ$ line indicating comparable linear and interaction effects. The rest of the coefficients generate significantly less perturbations in social welfare. In case of the financial result in the right hand side, travel time still appears to be an important factor, while the frequency related operational cost parameter ($v$) is now almost as relevant as the value of in-vehicle time. In this case most parameters have stronger nonlinear and interaction effects, suggesting that a more sophisticated variance decomposition method may be needed to measure the relative importance of our model parameters.

Sobol’ indices are derived from a functional decomposition of the variance of optimal social welfare and profit outputs (Sobol’, 1990). First order (often referred to as main effect) indices measure the share in the variance of the output that can be attributed to a given input parameter. For input $X_i$, the main effect Sobol’ index is given by $S_i = D_i(Y)/\text{var}(Y)$, where $D_i(Y) = \text{var}[E(Y|X_i)]$ and $Y$ is the output variable. The total effect index, introduced by Homma and Saltelli (1996), adds all additional interaction effects of the input to its first order index, thus measuring the global influence of each parameter on output sensitivity.
These variance based importance metrics can be estimated from a Monte Carlo sample with mutually independent input variables, with standard errors derived by bootstrapping. Figure 14 visualises the Sobol’ indices of seven input variables of our numerical model.

We estimated main effect and total effect Sobol’ indices using the Jansen estimator (Jansen, 1999, Saltelli et al., 2010) of the sensitivity package of R. Very similar results can be derived from the Martinez estimator. In case of the sensitivity of social welfare, travel time has a distinguished role, in line with the results we inferred with the Morris method. The value of in-vehicle time is again the second most important input. The variance of the financial result as model output, on the other hand, can be equally well explained with frequency-related parameters, $a$ and $v$. In Section 5 we focus more on the variables to which the output variables are more sensitive.

Based on randomly drawn Monte Carlo samples we also investigated whether the func-
tional relationships between the magnitude of demand imbalances and the volumes in Table 3 are robust with respect to the inputs. We assumed that all parameters are normally distributed\(^{17}\). The only exception is \(\delta\), which is drawn from a uniform distribution between 0.8 and 1, to avoid any misleading model outcomes due to diseconomies of scale in vehicle size. Under these assumptions all results in Table 3 hold except for the financial result (optimal subsidy) in the MS case. For weak density economies in vehicle size (typically when \(\delta > 0.925\)), the optimal subsidy may decrease with the magnitude of demand imbalance, just like in case of willingness to pay asymmetries.

**Elementary effect of parameters**

As part of the sensitivity analysis, we run a sequence of simulations by halving and doubling the baseline parameters of Table 1 individually, focusing on the most influential parameters. All results reported in Table 3 are robust in the extended range of parameter values in case of the sign of reaction as the magnitude of demand imbalances grows. When it comes to the comparison between market size and willingness to pay imbalances, however, we observe certain sensitivity with respect to the parameters chosen. In the baseline framework we found that peak crowding is more severe in the MS case, which implies that peak fares as well as the average crowding experience per passenger are also higher. Now, these results change in a very similar way to more crowding in the WTP case when we double either the value of uncrowded travel time (\(\beta\)) or the travel time itself (\(t\)), or we halve the original value of maximum willingness to pay (\(A_0\)). We attribute this sensitivity to the fact that in the MS setting peak demand becomes more elastic with respect to user costs as \(\theta\) grows, and therefore the operator has to offer more capacity (less crowding) to maximise welfare. When \(\beta\) and \(t\) are higher, user costs become more relevant, and therefore peak demand sensitivity may become more of an issue. The reduction in \(A_0\) also points in the direction of increased demand elasticity.

The model’s sensitivity to travel time \(t\) resembles what we observe when comparing short and long distance public transport: for higher travel time, the optimal frequency decreases and larger vehicles are operated, as one may infer from equations (17) of the analysis of inelastic demand. The optimal occupancy rate is inversely related to travel time, which is a result that we could not identify with inelastic demand. For long distance services in the WTP scenario the threshold value of \(\theta\) where the off-peak market is not served any more decreases; in case of \(t = 30\) min travel time it drops to \(\theta = 0.8\), from the original value of \(\theta = 0.88\) associated with \(t = 15\) min. All other findings summarised in Table 3 are robust with respect to travel time.

The magnitude of the user preference parameters, i.e. \(a\), \(\beta\) and \(\varphi\), are not expected to

\(^{17}\)We set the means to the original values provided in Table 1 and the standard deviations to 20% of the mean.
vary significantly in real applications compared to the assumed values, as these are in line with a wide body of empirical findings in the transport demand modelling literature. We assumed $\beta = 20$ monetary units for the uncrowded value of in-vehicle time, which is certainly realistic in British pounds, and not far from reality in case of Western European countries and the US. Waiting time is generally more inconvenient than travelling inside the vehicle. We multiplied $\beta$ by 1.5 to get the value of waiting time. This parameter has of course a huge impact on the trade-off between the optimal frequency and vehicle size, but the shape of the investigated relationships remains robust. The impact of the crowding multiplier ($\varphi$) is far less considerable on the optimal frequency, but it does affect vehicle size, crowding levels and the resulting peak and off-peak fares in the intuitively plausible way. It has no significant impact on our basic findings regarding the economic and financial implications of demand imbalances.

The original operational cost parameters ($v, w$) were selected from other simulation studies in the literature and supported by the realism of the resulting crowding patterns only, with no firm empirical evidence. Therefore it may be useful to estimate these parameters from real operational data as part of future research efforts. Unsurprisingly, halving or doubling $v$ has the opposite impact on frequency and vehicle to what we found for the waiting time coefficient. The model collapses for very high values of either $v$ or $w$, because if service provision is extremely expensive, having no passengers at all may maximise social welfare. Our results are robust with respect to the baseline market size ($M_0$) as well. Both frequency and vehicle size increase heavily as we double $M_0$ from 5 to 10 thousand passengers per hour, and due to the presence of scale economies in vehicle size ($\delta < 1$), the optimal average occupancy rate slightly decreases. Nevertheless, all other findings reported earlier in this section remain unaffected by the scale of operations.
References


Tirachini, A. (2014), ‘The economics and engineering of bus stops: Spacing, design and congestion’, 

light rail, heavy rail and bus rapid transit over a radial public transport network’, Research in 
transportation economics 29(1), 231–242.

Tirachini, A., Hensher, D. A. and Rose, J. M. (2014), ‘Multimodal pricing and optimal design of 
urban public transport: The interplay between traffic congestion and bus crowding’, Transportation 

Wardman, M. and Murphy, P. (2015), ‘Passengers’ valuations of train seating layout, position and 


Yang, H. and Meng, Q. (2002), ‘A note on “highway pricing and capacity choice in a road net-
work under a build-operate-transfer scheme”’, Transportation Research Part A: Policy and Practice 
36(7), 659–663.