Public transport provision under agglomeration economies

Daniel Hörcher∗1, Bruno De Borger2, Woubit Seifu2, and Daniel J. Graham1

1Imperial College London, United Kingdom
2University of Antwerp, Belgium

19 December 2019

Abstract

The purpose of this paper is to investigate, using both theoretical and numerical analysis, the impact of agglomeration externalities on short-run policy decisions in public transport, i.e. socially optimal pricing, frequency setting, and subsidisation. We develop a simple two-mode model in which commuters can opt for car or public transport use; car use leads to congestion, and public transport is subject to crowding. Allowing for agglomeration externalities, we show the following results. First, if car use is correctly priced for congestion, agglomeration benefits imply substantially lower public transport fares and higher frequencies. They neutralise to some extent the pressure to increase fares to correct the crowding externality. Second, as a consequence, agglomeration benefits justify low cost recovery ratios in public transport. Assuming an agglomeration elasticity of 1.04, a value well within the range of reported elasticities, numerical implementation of the model finds that cost-recovery ratios are 35.8% lower than in the absence of the productivity externality. Third, interestingly, the effect of agglomeration benefits on fares and frequency is much smaller if road use is exogenously under-priced. In this case, any modal shift induced by lower public transport fares has opposing agglomeration effects on the two modes, since agglomeration benefits are not mode-specific.

Keywords: agglomeration, public transport, congestion, crowding, subsidies

*Corresponding author – Address: Skempton Building, South Kensington Campus, London SW7 2AZ; Electronic address: d.horcher@imperial.ac.uk
1 Introduction

Since the appearance of advanced methods in the empirical measurement of agglomeration externalities, such ‘wider economic benefits’ have gained popularity in large-scale project appraisal (Laird and Venables, 2017, Graham and Gibbons, 2019). There is no reason to believe, however, that agglomeration benefits are less relevant for short-run supply optimisation. This has been recognised by Arnott (2007), who showed in a numerical example that congestion and agglomeration externalities can be of similar magnitude. The purpose of this paper is to investigate the impact of agglomeration externalities on socially optimal public transport fares and frequency. We will show that agglomeration benefits strongly affect optimal public transport policies. Moreover, we find that the effects on fares, frequencies and optimal public transport subsidies strongly depend on pricing policies for the substitute car mode and the degree of modal substitution.

The problem of determining the socially optimal fare and frequency for public transport operators was first convincingly analysed in the seminal papers of Mohring (1972, 1976). He pointed at the presence of economies of density in scheduled public transport services, due to the reduction in average waiting time when ridership and thus the optimal frequency increase. Under constant returns to scale in operational costs and random passenger arrivals to stations (so that the average waiting time equals half of the headway), the marginal cost of frequency adjustment is neutralised by the marginal benefit of waiting time reduction. As a consequence, marginal cost pricing in Mohring’s model leads to a zero fare, and operating costs have to be fully recovered by subsidies.

Of course, Mohring’s model has been extended in various dimensions, and his conclusions have been qualified. First, public transport in urban areas is often subject to crowding so that additional passengers impose an external cost on other users (Jansson, 1980, Jara-Díaz and Gschwender, 2003). Recent empirical evidence suggests that the crowding externality is numerically not negligible, so that crowding may substantially increase welfare-optimal fares\(^1\). Second, public transport policies cannot be evaluated without properly taking into account pricing policies for car use, the main substitute mode. If car use is priced inefficiently low, public transport subsidies can be defended as a standard second-best policy, substantially alleviating the social costs of congestion (see, for example, Parry and Small, 2009, Anderson, 2014). However, the slow but steady movement towards some form of road congestion pricing in a number of cities (Anas and Lindsey, 2011) reduces the importance of the second-best argument. Not surprisingly, recent papers that look at jointly optimal prices for road and public transport use in the presence of road congestion and crowding find much higher optimal public transport fares (see, for example, Basso and Silva, 2014, Kilani et al., 2014). Third, subsidies to public transport are socially costly if they have to be financed by distortionary taxes. A non-zero marginal cost of public funds suggests that large subsidies

\(^1\) On public transport pricing in the presence of crowding see, among others, Tirachini et al. (2010), De Borger and Proost (2015) and Höcher (2018). Numerical evidence on crowding is also provided by, for example, Tirachini et al. (2013), Haywood and Koning (2015) and Höcher et al. (2017).
may be inefficient\(^2\) (Proost and Van Dender, 2008). In sum, fares much below the social cost of providing public transport services may be inefficient, especially if road use is appropriately priced.

However, the previous literature ignores that the Mohring effect is not the only external benefit of public transport provision. In high density metropolitan areas public transport predominantly facilitates commuting to work. Cheap commuting induces labour supply, and this generates agglomeration benefits in the urban labour market (Safirova, 2002, Duranton and Puga, 2004, Rosenthal and Strange, 2004, VENABLES, 2007). The increase in the density of economic activities may raise productivity through several channels. Knowledge spillovers generate positive externalities in production and may stimulate faster adoption and penetration of new technologies. Moreover, a large labour market facilitates matching workers and firms; it therefore reduces the probability that workers will not find a job as well as the probability that firms will be unable to fill vacancies.

In the rest of this paper we develop a simple two-mode model in which transport users can opt for car or public transport use. Car use leads to congestion, and public transport is subject to crowding. Overall commuting generates agglomeration externalities. The model takes into account the cost of funds to finance the public sector deficit. Using a basic representation of public transport operations, we identify the role of agglomeration benefits in the rules describing socially optimal public transport fares and frequencies and, by implication, the effect of agglomeration on the cost recovery ratio of the public transport system. To emphasise how the results depend on road congestion for car traffic and on the pricing policies to deal with congestion externalities, we illustrate the theoretical model with a numerical application under three different scenarios. For pedagogic reasons, we first assume that public transport operates in isolation and does not face competition from car use. We then introduce car use but assume that drivers pay a fixed (inefficient) fuel tax, i.e., road congestion is unpriced. Finally, we reconsider optimal fares and frequencies under agglomeration economies when road congestion is socially optimally priced. The numerical model uses recent empirical estimates of (i) operational cost elasticities, (ii) agglomeration externalities and (iii) crowding cost functions, and it derives numerical results on optimal fares, frequencies and subsidy levels in a representative urban area.

As far as we know, this is the first paper that explicitly considers the potential role of agglomeration benefits for public transport policies. It studies a tension of positive externalities (agglomeration benefits, economies of density) and negative externalities (crowding and congestion). This tension also underlies the literature on the optimal size of cities (Henderson, 1974) and the new economic geography (Fujita and Krugman, 2004). The paper contributes to the literature with three key findings. First, we theoretically consider the effect of urban agglomeration productivity on optimal fares and frequencies for public transport. We extend existing expressions for optimal fares and frequency to capture agglomeration benefits, showing that they neutralise to some extent the negative consumption externalities (e.g.,

\(^2\)Of course, there are arguments for low public transport fares for other than efficiency reasons. It is often argued that fares should be low for income distributional reasons. The relevance of this argument depends on the extent to which lower income classes are more intensive public transport users than the rich.
crowding) that might otherwise justify high socially optimal public transport fares. Second, we develop a numerical version of the model and show that agglomeration economies have a crucial impact on the optimal fare-frequency combination — and on the implied subsidy rate — when car use is optimally priced, but that this effect is reduced if road use is under-priced. Intuitively, agglomeration externalities are not mode-specific so that, if the productivity externality cannot be internalised on the road, then any mode shift induced by lower public transport fares has opposing agglomeration effects in the two modes. This also implies, with a reference to Arnott (2007), that reducing road taxes in light of potential agglomeration benefits may have weak impact if public transport remains sub-optimally priced. Third, we find that under optimal road pricing the effect of agglomeration on public transport pricing and subsidies is comparable in magnitude to that of other crucial determinants, such as economies of density and crowding. One interpretation is that high agglomeration benefits ‘justify’ low cost recovery ratios in public transport. Further note that transport demand, crowding in public transport, and wages are much less sensitive to agglomeration, from which we infer that sub-optimal subsidisation might not be directly visible in reality.

An overview of the paper is as follows. In Section 2 we discuss the relation of the current paper with the earlier literature. Section 3 presents the theoretical model and interprets the analytical results we derive for the optimal fare and frequency. In Section 4 we discuss the numerical implementation of the model. Numerical results are presented in Section 5. We discuss the implications of agglomeration benefits for fares, frequencies, road tolls and wages. Moreover, we report on the sensitivity of the results with respect to key input parameters, including the agglomeration elasticity, the importance of crowding, and the cost of public funds. In Section 6 we zoom in on the implications of agglomeration for the cost recovery ratios and optimal subsidies in the public transport sector. Section 7 concludes.

2 Previous literature

In this section we briefly review the link between the current paper and the existing literature. Our model relates to at least four types of studies.

First, since the seminal work of Mohring (1972, 1976), an extensive literature focused on improving the representation of public transport user costs, operational costs, and the demand side of the transport market. As for user costs, it was recognised that on-vehicle

3A series of studies further analysed how the cycle time of a service can be affected by boarding and alighting passengers (Jansson, 1980, Jara-Díaz and Gschwender, 2003, Jara-Díaz and Gschwender, 2009, Tirachini et al., 2010, 2014). Boarding and alighting passengers impose travel time losses on those already on board, so that the magnitude of the crowding externality increases in the load factor at the moment of boarding. When the comfort level of standing and seated users are differentiated, and since using a seat imposes an occupancy externality on standees, Kraus et al. (1991) found crowding costs to decline with the load factor. One implication would be that passengers boarding earlier should pay higher fares to internalise the external cost (Hörcher et al., 2018). Most recently, it has been argued that crowding costs are most realistically seen as piecewise linear, with two clearly identifiable kinks: one when seating capacity is reached, and another one when buses are full (De Palma et al., 2015). However, recent empirical crowding models estimated for Paris find that linear crowding functions describe the data quite well (Haywood and Koning, 2015).
crowding imposes an externality on other users. Crowding was first captured by imposing an explicit capacity constraint on vehicles (Jansson, 1980, 1984); later models incorporated crowding as a determinant of passenger comfort, resulting in a value of in-vehicle time that depends on the number of passengers on board of the vehicle (see, for example, Jara-Díaz and Gschwender, 2003). Operational costs of public transport operators were refined in a number of extensions as well, mainly focusing on bus services. Apart from fares and frequency, studies have analysed fleet size, i.e. the optimal number of vehicles (De Borger and Wouters, 1998, Jara-Díaz et al., 2017), optimal vehicle size (Oldfield and Bly, 1988, Jara-Díaz and Gschwender, 2003, Hörcher and Graham, 2018, Börjesson et al., 2018), and the number of stops (Tirachini et al., 2010, Tirachini, 2014). Moreover, although some early models considered a single bus line with exogenous demand (Jansson, 1980), others introduced price-responsive demand and explicitly distinguished peak and off peak periods (Chang and Schonfeld, 1991, Jansson et al., 2015, Jara-Díaz et al., 2017). More recently, Hörcher and Graham (2018) derived general results for multi-period supply.

Second, our model relates to the literature integrating the transport, labour and housing markets. A series of models make the link between commuting and labour supply explicit and jointly study transport prices and taxes on labour (Mayeres and Proost, 1997, Parry and Bento, 2001, De Borger and Van Dender, 2003, De Borger and Wuyts, 2011, Tikoudis et al., 2015). The public transport component of these models is typically highly stylised; they do not optimise frequency, and they neither include agglomeration benefits nor crowding externalities.

Third, we contribute to the literature on agglomeration economies. Productivity in the urban economy increases with total labour supply, i.e. the effective economic density of a region, as a result of sharing public goods and inputs, improved matching between various agents, and learning through knowledge spillovers (Graham, 2007a, Greenstone et al., 2010, Puga, 2010). Provided that transport improvements contribute to the effective density of a city, they lead to higher wages and increase households’ individual labour supply (Venables, 2007). Agglomeration benefits are now routinely accounted for in cost-benefit appraisals of large-scale infrastructure projects (Mackie et al., 2011, Laird and Venables, 2017, Graham and Gibbons, 2019), as well as in model of congestion pricing (Safirova, 2002, Arnott, 2007, Zhang and Kockelman, 2016, Fosgerau and Small, 2017). However, their implications for pricing and frequency of public transport operations remain to be analysed.

4Public transport is equally stylised in models with a formal land market and endogenous residential location and city size. For example, Brueckner (2005) studies the impact of transport costs on households’ demand for residential space. He shows that sub-optimal subsidies for commuting might lead to inefficient urban sprawl, assuming that the transport system features constant returns to scale. To the extent that agglomeration benefits exist optimal subsidies will be larger than absent this positive externality. However, as the mechanisms that drive Brueckner’s result will still be operational once one allows for agglomeration benefits, his overall message that inefficiently high subsidies may lead to urban sprawl is likely to remain valid.

5The productivity effect can be measured through agglomeration elasticities, i.e. the elasticity of output with respect to total labour supply. In many empirical analyses total labour supply is replaced with the effective economic density of a region, defined by the availability of work force in neighbouring regions, discounted by a distance decay effect directly related to the generalised cost of travelling between the two regions (Graham, 2007a,b). See the meta-analyses of Melo et al. (2009) and, more recently, Maré and Graham (2013).
Finally, the current paper is related to the literature on optimal public transport subsidies. In the absence of first-best pricing on congestible roads, reducing public transport fares below marginal social cost can be a legitimate second-best policy. The general idea that public transport can serve as a tool to alleviate the deadweight loss of sub-optimally priced road congestion is by now widely documented. De Borger and Wouters (1998) and Proost and Van Dender (2008) found low optimal fares and high subsidies for Belgium and the city of Brussels, respectively. Similarly, Parry and Small (2009) show that in 2001 the optimal second-best subsidy for most forms of public transport in large agglomerations like London was higher than the subsidy actually observed\(^6\). Other recent studies supporting the above idea include Anderson (2014) and van Ommeren et al. (2018). However, when road use is correctly priced and there is important crowding in public transport, the need for large subsidies is much less clear. Basso and Silva (2014) looked into a wider set of policy interventions than subsidies for public transport: they also study congestion pricing for car use, the role of dedicated bus lanes and the effect of peak differentiation for bus fares. They find, also for London, that congestion pricing and dedicated bus lanes (absent bus subsidies) are much more efficient than bus subsidies. Kilani et al. (2014) studied a joint pricing reform of car and public transport use in Paris, taking into account that both the road and public transport networks are highly congested. Contrary to earlier studies, they found that even at current (too low) road charges, it is best to increase peak period prices of public transport on some highly crowded links. Moreover, fares should be increased in general whenever road pricing is put in place. Finally, Börjesson et al. (2018) studied the Stockholm situation after the cordon toll. They find that the best overall transport policy reform includes, among others, higher toll levels on car use, higher peak bus fares and frequency, and free off-peak bus services offered at lower frequencies. These policies do not necessarily mean higher bus subsidies. Moreover, since optimal pricing is found to only marginally improve welfare compared to optimising frequencies, they argue that one may as well start by optimising frequency.

Compared to the literature surveyed above, our model uses a simplified representation of public transport operations, focusing mainly on fares and frequencies, ignoring other decision variables such as the number of stops and vehicle size. However, our model does capture agglomeration benefits, it distinguishes two competing modes, it includes road congestion as well as crowding externalities, and it explicitly considers the role of correctly pricing car use for the results.

### 3 Theoretical analysis

In this section we introduce an analytical framework from which we derive analytical results for the optimal fare and frequency of a public transport service, taking the representative

---

\(^6\)The main intuition for the very high optimal subsidy they obtained was not only important unpriced road congestion, but also the (probably overly optimistic) assumption that half of the attracted public transport users were former car users. Another assumption underlying their model was that the frequency of public transport was related to demand by a simple rule of thumb, rather than explicitly optimised. The optimal subsidy they obtained was in some cases of the order of 90\% of operating costs.
household’s endogenous labour supply decision and urban agglomeration economies into account. The resulting supply rules are aimed to be comparable with earlier analytical formulae in the literature.

3.1 Household behaviour

Consider an urban agglomeration with \( N \) residents assumed to be homogeneous, so that we can focus on the behaviour of a representative household. It derives utility from a quasiconcave function of consumption \( y \) and leisure time \( l \). The unit price of consumption goods is normalised to one. The household’s income is determined by the intensity of its individual labour supply. We assume perfect complementarity between labour supply and commuting, so that the number of active workdays within a period of time equals the number of commuting roundtrips. We denote the total number of roundtrips per time period a consumer makes by \( x \). Trips can be made either by public transport \( (x_r) \) or by automobile \( (x_c) \). Therefore, individual labour supply is \( x = x_r + x_c \). Public transport operates on isolated infrastructure so that there are no interactions with private car use. In other words, the model focuses on a rail service (or a bus service running on a separate bus lane), but it can be extended into a setting with shared road use.

Admittedly, although not uncommon in the literature, the assumption of perfect complementarity between labour supply and commuting is quite restrictive. There are three reasons why commuting and labour supply may deviate from one another: (i) commuters may work more hours per day but fewer days per week. Then labour supply, measured in hours per week, may remain constant or even increase while the number of commuting trips declines; (ii) travel demand may increase while labour supply does not, due to non-commuting trips (e.g. leisure or shopping trips); (iii) telecommuting and other types of remote working may imply that aggregate labour supply increases but commuting demand does not.

Although all three hypotheses are theoretically relevant, for the purpose of this paper we kept the assumption of perfect complementarity. First, the potential substitution between days per week and hours per day does not seem to be empirically convincing. For example, Gutiérrez-i-Puigarnau and van Ommeren (2010) find that the sample of commuters who face an exogenous change in commuting costs but do not relocate, do not change the number of workdays per week either, and the impact on daily working hours is also negligible. In other words, substitution between weekly workdays and daily working hours is limited. Second, little is known about the contribution of various forms of teleworking on productivity, both at the firm level and in terms of agglomeration effects. In general, insufficient reliable information is available to capture the effect of agglomeration in a model with multiple

---

7 We adopt the definition of Anderson et al. (1992) who state that “a representative consumer is a fictitious agent whose utility embodies aggregate preference for diversity. […] The representative consumer plays two roles. First, the demand system for the differentiated product is generated by maximizing the representative consumer’s utility function. Second, welfare judgements are based explicitly on this utility function”.

8 Recent research somewhat qualifies this conclusion, however. Although traditionally empirical evidence on the relationship between commuting costs and aggregate labour supply was scarce (see Gibbons and Machin, 2006), recently Gershenson (2013) and Fu and Viard (2018) report a negative causal relation.
modes, multiple trip purposes and the possibility of telecommuting. Third, note that the potential bias in the effects of agglomeration on public transport fares and frequencies due to telecommuting and non-commuting traffic work in opposite directions, so that they partly cancel out.

The net income earned after a workday is the difference between the daily wage \( w \) and the monetary price of travelling in each mode. Rail users pay a fare \( p \) after each trip; car users’ expenditures include a monetary operational cost \( c \) and, if levied by the government, a road tax \( \tau \). The household’s budget constraint is given by

\[
y + px_r + (c + \tau)x_c = w(x_r + x_c).
\]

We keep the length of a workday constant at \( L \) hours per day. The commuter’s time constraint can be written as

\[
(L + t_r)x_r + (L + t_c)x_c + l = \bar{L},
\]

where \( t_r \) and \( t_c \) are travel times by rail and car, and \( \bar{L} \) is total time available.

Rail travel time is assumed to depend on the frequency of service offered. It is specified as \( t_r = t_v + 0.5f^{-1} \), where \( t_v \) is a fixed in-vehicle time and \( 0.5f^{-1} \) is half of the headway, i.e., the average waiting time denoting service frequency with \( f \) and assuming random arrivals to the origin station. We normalise \( t_v \) to zero without loss of generality. Road congestion is captured by a simple linear congestion function relating travel time to the total number of trips: \( t_c = \alpha N x_c \). We consider a given road network; the slope of the congestion function remains constant throughout the analysis.

Public transport passengers not only spend time when making trips, they also potentially suffer from other inconveniences, such as crowding. This causes disutility, directly captured by an additively separable sub-utility function. More precisely, the household’s utility function is specified as

\[
u = U(y, l) - x_r \cdot \varphi(N x_r f^{-1}).
\]

In this expression, \( \varphi \) reflects the inconvenience associated with rail trips, expressed as a function of the average occupancy rate \( N x_r f^{-1} \). Note that we neglect the household’s potential scheduling behaviour, and model the morning and afternoon peaks in a static framework in which demand is homogeneously distributed. The occupancy rate is assumed to be exogenous to the individual decision maker, so the crowding externality that passengers impose on each other is not internalised without the supplier’s intervention.

---

9De Borger and Wuyts (2011) present an analytical model with telecommuting and two trip purposes. They assume that telecommuting is less productive than working at the office, but it saves on commuting costs. However, they do not include agglomeration effects.

10Throughout this paper we consider an urban mass transport system, assuming that the rail service has a fixed cycle time as long as service frequency is below the line’s capacity. The analysis could be extended for the case (more appropriate for some urban bus services that interact with car traffic) where frequency affects travel times through congestion, and dwell times would be sensitive to the number of passengers boarding at the front door.
Users maximise utility (3) by selecting the optimal level of leisure, consumption, and commuting demand by rail and car, subject to the monetary and time constraints (1) and (2). The corresponding Lagrangian can be written as

$$\Lambda_u = U(y,l) - x_r \cdot \varphi(\cdot) + \kappa [(w - p)x_r + (w - c - \tau)x_c - y] + \mu [\bar{L} - (L + t_r)x_r - (L + t_c)x_c - l],$$

where $\kappa$ and $\mu$ are Lagrange multipliers denoting the shadow price of money and time, respectively. From the first-order conditions of (4), one obtains the marginal value of leisure time $VoT$ at the optimum, expressed in monetary terms. It is easy to show that

$$VoT = \frac{\mu}{\kappa} = \frac{w - p}{L + t_r} - \frac{\varphi \kappa^{-1}}{L + t_r} = \frac{w - c - \tau}{L + t_c}.$$

As given by (5), the value of time can be expressed as the wage net of the rail fare per hour working and commuting by rail, corrected for the effect of crowding disutility. Note that in the present setup, the net real wage for rail users is greater than for car commuters, due to the inconvenience of crowding.

The solution of the utility maximisation problem leads to the indirect utility function

$$V(p, f, \tau; t_r, t_c, Nx_r f^{-1}, w).$$

Indirect utility is a function of car and rail prices and public transport frequency, conditional on all factors exogenous to the individual’s decision; these include travel times, the level of crowding, and the wage. To simplify notation, in what follows we suppress these arguments of $V(p, f, \tau)$, focusing on the three decision variables set by the authorities.

### 3.2 Firm behaviour with agglomeration benefits

On the production side of the economy we assume that firms in perfect competition produce a homogeneous output using only labour as an input. The output of representative firm $j$ is determined as

$$q_j = A(Nx) \cdot I_j,$$

where $I_j$ denotes its labour input, and $A(Nx)$ measures productivity in function of total labour supply in the economy; $A_x > 0$. That is, individual firms face constant returns to scale and consider productivity as an exogenous parameter, while the economy as a whole features increasing returns to scale. Firms maximise profits in this perfectly competitive environment, so that the equilibrium wage becomes the internal marginal product of labour.

With this specification, we follow the agglomeration literature, which provides convincing evidence of the existence of agglomeration benefits, and reports a range of estimates of the
effect of labour supply on productivity (see the review articles by Graham and Gibbons, 2019, Melo et al., 2013). This literature typically measures productivity in function of Access To Economic Mass (ATEM), whereby in almost all recent studies the ‘mass’ variable is assumed to be employment. It is worth emphasising that it is not an increase in labour supply per se that generates the productivity benefits, but rather the associated externalities from matching, sharing and learning in the enlarged local market.

In subsequent numerical analyses we specify the productivity term as

\[ A = a \cdot (Nx)^{\eta-1}, \]

where \( \eta \) is the elasticity of output with respect to total labour supply. Thus, the equilibrium wage becomes

\[ w = A(Nx) = a \cdot (Nx)^{\eta-1}. \]  

(8)

We assume the commuter’s individual labour supply has no sensible effect on her own wage. However, the external effect of working on the productivity of the population as a whole may not be negligible, and therefore we do consider this agglomeration externality in the derivation of welfare maximising public transport supply. Note that we assume here that agglomeration is independent of the mode of commuting: \( x_r \) and \( x_c \) have identical contributions to urban productivity.

3.3 The social optimum

Suppose a social planner determines transport prices and public transport frequency; the planner cares about the well-being of the representative consumer and the net revenues of the transport system. These are defined as the revenues from public transport fares and road tolls minus the operational costs of public transport provision. The latter are captured by a function \( z(f) \) of frequency, where \( z' > 0 \). This specification implies that the costs associated with carrying passengers are ignored. However, as in the numerical implementation of the model, optimal frequencies yield occupancy rates not too close to capacity, the marginal operating cost of extra passengers is very small (see the references given in the literature survey above). The social planner is assumed to solve the following problem:

\[ \Lambda_s = N \cdot V(p, f, \tau) + \beta [N(x_r p + x_c \tau) - z(f)] \]

(9)

where \( \beta \) reflects the relative value the government assigns to net transport revenues. As noted by Proost and Van Dender (2008), it can be directly related to the marginal cost of funds, see below\(^{12}\).

We can write the first-order condition with respect to the optimal public transport fare

\(^{12}\)Alternatively, with very minor modification (i.e. adding an exogenous level of net revenues to be generated), expression (9) can be interpreted as maximising welfare subject to a government budget restriction. The parameter \( \beta \) is then the Lagrange multiplier associated with the constraint. We prefer the interpretation in the text, because the numerical model below will use an exogenous value of the tax revenue premium implied by \( \beta \), derived from figures available in the literature.
in the following form (see Appendix A):

\[
p^* = \frac{1}{\beta} x_r \varphi' N f^{-1} - \frac{\kappa}{\beta} x w' N \frac{\beta}{\alpha_p} + \frac{\kappa - \beta}{\beta} x_r + \left[ \frac{\mu}{\beta} x_c t_c' N - \frac{\kappa}{\beta} x w' N - \tau \right] \frac{\partial x_c}{\partial p} \frac{\partial \tau}{\partial x_c}.
\]  

(10)

In this expression \( \kappa \) and \( \mu \) are the marginal utilities of money and time defined before, see the Lagrangian in (4). Of course, (10) is not a closed-form solution, as the right-hand side depends on all endogenous variables, but it is useful to interpret the determinants of the optimal fare, conditional on given road tolls and public transport frequency. It consists of four components. The first term on the right-hand side reflects the marginal external crowding inconvenience imposed on fellow users. It is expressed in monetary terms by dividing by the value of money to the government. The second term is the marginal agglomeration benefit of public transport use: more commuting generates a positive productivity externality, captured by increasing wages. The third component of (10) reflects a revenue generating mark-up; it is positive provided the government’s valuation of extra revenues exceeds the consumer’s valuation of income, i.e. \( \kappa < \beta \). For \( \kappa = \beta \), the markup is zero. Note that it also depends on the sensitivity of rail demand: the more \( x_r \) is insensitive to the fare, the more extra revenues raising the fare will generate. Finally, the fourth component relates to pricing of the competing mode. The terms between the square brackets capture the degree to which the road toll takes into account the marginal cost of congestion and the marginal benefit due to agglomeration. If the road toll falls short of the marginal external congestion cost corrected for the agglomeration benefit, road traffic is under-priced; this leads to a reduction in the socially optimal fare, a standard second-best argument. The size of the reduction depends on the degree of substitution between the two modes (see the term \( \frac{\partial x_c}{\partial p} / \frac{\partial x_c}{\partial \tau} \), which is negative if car and rail use are substitutes).

The introduction of agglomeration in the determination of the socially optimal public transport fare leads to two relevant questions. First, from a policy point of view what matters is the relative magnitude of the first two terms on the right hand side of (10), the crowding externality and the agglomeration benefit. The crowding externality is certainly non-negligible, especially in peak periods (Haywood and Koning, 2015). We assess the interplay between the crowding and agglomeration in the numerical application below. Second, what is the overall sign of the agglomeration benefit, knowing that it operates both through commuting by public transport (the second term) and by the private car (the fourth term). Note that \( \partial x_c / \partial p > 0 \) and \( \partial x_p / \partial p < 0 \); however, the cross-effect is normally smaller in magnitude due to imperfect substitution, so that one expects the ratio \( \frac{\partial x_c}{\partial \tau} / \frac{\partial x_c}{\partial \tau} \) to be between \(-1\) and zero. That is, the net agglomeration effect of the two modes jointly is expected to be negative under imperfect substitution.

Completely analogous, we find the following expression for the socially optimal road toll, conditional on a given fare and public transport frequency (see again Appendix A):

\[
\tau^* = \frac{\mu}{\beta} x_r t_c' N f^{-1} - \frac{\kappa}{\beta} x w' N \frac{\beta}{\alpha_r} + \frac{1}{\beta} x_r \varphi' N f^{-1} - \frac{\kappa}{\beta} x w' N - p \frac{\partial x_c}{\partial \tau} \frac{\partial \tau}{\partial x_r}.
\]  

(11)
Interpretation is analogous to that of (10).

Finally, turn to the optimal frequency of the public transport service for arbitrary prices for road and rail use. The optimal frequency rule takes the following form (see Appendix A):

\[ f^* = \sqrt{\frac{\mu \beta^{-1} 0.5 x_r}{\frac{1}{\beta} x_r \varphi_f + \frac{\varphi'}{N} + \left[ \frac{1}{\beta} x_r \varphi_{x_r} - \frac{\kappa}{\beta} xu'N - \tau \right] \frac{\partial x_r}{\partial f} + \left[ \frac{\kappa}{\beta} xu'N - \tau \right] \frac{\partial x_r}{\partial f}}}. \] (12)

The optimal frequency rule follows the square root principle of Mohring (1972): the optimal frequency increases with the square root of rail demand. The marginal utility of leisure time ($\mu$) has a positive impact on frequency through passengers’ valuation of frequency-dependent average waiting times, also in the numerator. The denominator has four components. First, frequency increases when it is successful in reducing crowding. This is captured by the first term: frequency alleviates crowding, so that $\varphi_f < 0$; therefore, the greater this marginal impact in absolute value, the higher the optimal frequency. The second element is the marginal operational cost of increasing frequency to the government (hence, evaluated at the social value of transport net revenues). The two remaining terms in the denominator reflect the welfare effect of demand induced by a marginal change in frequency in both modes. More frequency raises rail demand (see the third term in the denominator). If rail use is priced below its (net of agglomeration benefits) marginal crowding cost, then this is an argument for lowering frequency as well. Sub-optimally low fares imply overconsumption, and low frequency is used to discourage the marginal user from making a wasteful trip. The opposite argument applies for fares that are too high. The last term in the denominator reflects the idea that, if road tolls are below (net of agglomeration benefits) marginal external congestion cost and $xc$ is sensitive to rail service quality, then it is socially optimal to attract drivers off the road with higher frequency.

Joint optimisation of the two prices and public transport frequency can be shown – combining (10), (11) and (12) – to give the following socially optimal rules:

\[ p^*_1 = \frac{1}{\beta} x_r \varphi_f N f^{-1} - \frac{\kappa}{\beta} xu'N + \frac{\kappa - \beta}{\beta} x_1, \]
\[ \tau^*_1 = \frac{\mu}{\beta} xc't'N - \frac{\kappa}{\beta} xu'N + \frac{\kappa - \beta}{\beta} x_2, \]
\[ f^*(p^*_1, \tau^*_1) = \sqrt{\frac{\mu \beta^{-1} 0.5 x_r}{\frac{1}{\beta} x_r \varphi_f + \frac{\varphi'}{N} - \frac{\kappa - \beta}{\beta} \left( \frac{\partial x_r}{\partial f} + \chi_1 \frac{\partial x_r}{\partial f} + \chi_2 \frac{\partial x_r}{\partial f} \right)}}}. \] (13)

\(^{13}\) In its simplest form, the Mohring effect can be illustrated as follows. Assume the operator intends to minimise the sum of (i) the waiting time of users, $z_1 f^{-1} x_r$, which is linear in the headway between services ($f^{-1}$), and (ii) a linear operational cost function, $z_2 f$. The optimal frequency that minimises the total cost function $z_1 f^{-1} x_r + z_2 f$ is $f^* = \sqrt{z_1 z_2^{-1} x_r}$. That is, the optimal frequency is proportional to the square root of demand. This result reveals that the presence of waiting time benefits implies density economies in public transport provision, even if its technology exhibits constant returns to scale. Mohring’s result is preserved in the present model setup as well.
In these optimality rules $\chi_1$ and $\chi_2$ are functions of own- and cross-prices elasticities of demand in the two modes\textsuperscript{14}. The only reason why modal substitution may still matter under first-best road tolls is the presence of non-zero cost of public funds, reflected by $\kappa \neq \beta$. In that case, the socially optimal way of generating additional revenues depends to the price elasticity of demand in the two modes, reflecting the classical case of Ramsey pricing. Also note, consistent with the literature, that under optimal pricing the effects of induced demand do not appear in capacity (frequency) formulae (see e.g. Small and Verhoef, 2007, Section 5.1.1), because the envelope theorem ensures that induced demand has no welfare effect at the margin.

Finally, the first best can be described by a situation where the government has lump sum transfers available, so that $\kappa = \beta$. We then have

$$
p(f^*,\tau^*) = \frac{1}{\beta} x_r \varphi' N f^{-1} - x_w' N,
$$

$$
\tau(f^*,\tau^*) = \frac{\mu}{\beta} x_c t_c' N - x_w' N,
$$

$$
f(f^*,\tau^*) = \sqrt{\frac{\mu \beta^{-1} 0.5 x_r}{\beta^{-1} x_r \varphi_f' + z'(f) N^{-1}}}.
$$

(14)

Obviously, in first-best fares and tolls should now just capture the marginal external costs of crowding (fare) and congestion (toll), and the marginal external benefit due to agglomeration.

4 Implementing the model

The previous section showed that, ceteris paribus, agglomeration economies imply that commuting should be encouraged by lower public transport fares and higher frequencies. This raises a number of further questions. First, as mentioned above, does the presence of agglomeration economies outweigh the argument for higher fares due to crowding externalities in public transport? Second, how does the pricing of road use affect the role of agglomeration benefits for public transport fares and frequency? Third, how does agglomeration affect the socially optimal cost recovery ratio of public transport operations?

To answer these questions, we perform numerical simulation exercises in three scenarios that differ in how we treat the competing car mode:

\textsuperscript{14}In particular,\

\[
\chi_1 = p \left( \varepsilon_r - \frac{\varepsilon_{rc} \varepsilon_{cr}}{\varepsilon_r} \right)^{-1} + \tau \left( \varepsilon_{rc} - \frac{\varepsilon_r \varepsilon_c}{\varepsilon_{rc}} \right)^{-1} \frac{x_c}{x_r},
\]

and

\[
\chi_2 = \tau \left( \varepsilon_c - \frac{\varepsilon_{rc} \varepsilon_{cr}}{\varepsilon_r} \right)^{-1} + p \left( \varepsilon_{cr} - \frac{\varepsilon_r \varepsilon_c}{\varepsilon_{rc}} \right)^{-1} \frac{x_r}{x_c},
\]

where $\varepsilon_r$, $\varepsilon_c$ are own-price elasticities of rail and car demand, and $\varepsilon_{rc}$ and $\varepsilon_{cr}$ are cross elasticities of rail and car demand with respect to the monetary price of the other mode, respectively.
A. Isolated public transport service with no modal substitution. We determine optimal public transport fare and frequency.

B. The two modes are imperfect substitutes. We determine optimal public transport fare and frequency for an exogenously given price of car transport.

C. Rail and car are imperfect substitutes. We determine socially optimal fares and frequencies and optimal road tolls.

Let us now discuss the numerical implementation of the theoretical model. This required some relatively minor modifications. The first challenge is to find a suitable functional form – i.e., a specification that is reasonably realistic while allowing enough behavioural flexibility – for the utility function in equation (3). After considering several alternatives, \( U(y, l) \) is specified as a concave quadratic relation:

\[
U(y, l) = \alpha y \cdot y + \alpha l \cdot l + 0.5 \beta y \cdot y^2 + \beta l \cdot l^2 + \gamma \cdot yl, \quad \beta_y < 0, \beta_l < 0. \quad (15)
\]

We restrict \( \beta_y \) and \( \beta_l \) to be negative in order to ensure that first-order conditions lead to a global optimum. Moreover, as an extension of Parry and Bento (2001), for the numerical work we reformulate the sub-utility function \( \varphi(T_r, T_c) \) to capture more than just crowding disutility. Specifically, utility is formulated as \( u = U(y, l) - \varphi(T_r, T_c) \), where the sub-utility function is also specified quadratically:

\[
\varphi(T_r, T_c) = \alpha_r \cdot T_r + \alpha_c T_c + 0.5 \beta_r T_r^2 + \beta_c T_c^2 + \gamma T_r T_c, \quad (16)
\]

In this expression, \( T_c \) and \( T_r \) are the total travel times the consumer spends in car and rail trips. The travel time by car depends on free-flow travel time \( t_f \), road capacity \( K \), and aggregate demand for road use \( N x_c \), according to the standard BPR (US Bureau of Public Roads) congestion function. That is,

\[
T_c = 2x_c \cdot t_c = 2x_c \cdot t_f \left[ 1 + 0.15 \left( N x_c K^{-1} \right)^4 \right]. \quad (17)
\]

Total perceived rail travel time \( T_r \) is defined as the average generalised travel time \( (g_r) \) multiplied by the number of rail trips, where \( g_r \) is the sum of waiting time and in-vehicle time, the latter augmented by a crowding dependent multiplier:

\[
T_r = 2x_r \cdot g_r = 2x_r \cdot \left[ t_v \left( 1 + \delta \phi \right) + \delta w 0.5 f^{-1} \right]. \quad (18)
\]

In this expression \( \phi \) is the average occupancy rate of rail vehicles. Later on we specify this variable as \( \phi = N x_r / (f h s) \), where \( h \) is the length of the morning and afternoon peak periods in hours, and \( s \) denotes vehicle size, i.e. the in-vehicle area available for passengers. Thus, the crowding-dependent multiplier in \( T_r \) is linear in the occupancy rate with slope \( \delta \), in line with the empirical literature (Wardman and Whelan, 2011).

\[\text{15}\] The driving time dependent component of \( \varphi \) might represent the additional disutility of driving on top of the pure disutility of lost leisure time captured in \( u(y, l) \). During the initial attempts of model calibration we tested simpler specifications for \( \varphi(T_r, T_c) \), including the one used in the analytical part of this paper. However, simpler models were unable to reproduce the pattern of demand elasticities we borrowed from the literature, as opposed to the quadratic specification in equation (16).
To determine a reasonable value for $\beta$, note that it captures the social value of extra net transport tax revenues. The social value of generating an extra dollar may exceed one dollar because the additional revenue allows reducing the most distortionary taxes (typically labour taxes) elsewhere in the economy. In a partial equilibrium framework, this has often been captured by associating a monetary value to a dollar transport tax revenue that exceeds one. Intuitively, valuing transport tax revenue at more than its face value reflects the benefit of allowing to reduce distortionary taxes in the economy while keeping the budget unaffected. Our formulation of the government’s objective function implies that $\beta$ can be directly related to an assumed marginal cost of public funds: $\beta = \kappa \cdot \text{MCPF}$, where, see above, $\kappa = \partial U / \partial y$ is the marginal utility of private income. The social value of tax revenues to the government exceeds the marginal utility of income to consumers to the extent that the cost of public funds exceeds one. To implement the model, we determine $\beta$ so as to reflect a tax revenue premium of 6.6% (as in Proost and Van Dender, 2008). In other words, we assume the social value of a dollar of transport tax revenue is $1.066.

To take into account that each working day in practice requires a roundtrip we adapt the monetary and time budget constraints accordingly:

$$y = (w - 2p)x_r + (w - 2c - 2\tau)x_c,$$
$$l = \bar{L} - (L + 2t_r)x_r - (L + 2t_c)x_c. \quad (19)$$

Based on these constraints, in Appendix B we derive the system of first-order conditions that determine the equilibrium levels of commuting demand in function of prices, the rail frequency, travel times and other exogenous parameters.

Finally, for the operational cost function $z(f)$ we assume a constant elasticity of operational costs with respect to frequency. Based on recent empirical estimates\textsuperscript{16} by Anupriya et al. (2019), we set the elasticity to $\omega = 0.6$. Moreover, we assume in the baseline scenario that the cost recovery ratio is $\rho = 0.6$, so that fare revenues cover 60% of operating expenditures, in line with industrial averages of European and North American urban rail providers. Noting that this creates a direct relationship between the scale parameter of $z(f) = z_0 \cdot f^\omega$ and rail fare revenues in the initial equilibrium, as by definition $\rho = N2x_r p / z(f)$, we can calibrate the missing scale parameter as:

$$z_0 = \frac{N2x_r p}{\rho f^\omega}. \quad (20)$$

In what follows, we discuss the calibration and numerical optimisation of our model with respect to supply-side decision variables. Note that we focus throughout on the two-mode model used to study scenarios B and C. The model version to analyse scenario A is easily established by normalising $\alpha_c, \beta_c$ and $\gamma_t$ in equation (16) to zero, and removing all other car use related components from the model.

\textsuperscript{16}This estimate is based on a unique panel dataset covering the largest metro networks of the globe. Its data items have been collected for almost two decades within public transport benchmarking groups facilitated by the Transport Strategy Centre (TSC) at Imperial College London. The econometric strategy of Anupriya et al. (2019) is also unique in the sense that it controls for the endogeneity of input prices in the underlying cost function specification.
Turning to the calibration of the model, all three scenarios of the numerical experiment are calibrated to comply with the initial equilibrium consistent with the demand levels and demand elasticities given in Table 1. The majority of model parameters were selected to reflect a representative rather than a specific urban public transport market. More specifically, the aim of the calibration exercise is to find the set of utility parameters in (15) and (16) with which the initial demand levels of Table 1 are recovered, and equilibrium demand reacts to price and frequency changes in line with the assumed set of elasticities. Note that the calibration is based on own- and cross-price elasticities of rail and car commuting with respect to the price and quality of the public transport service. Given our assumption of perfect complementarity between commuting and labour supply, one easily shows that the elasticity of total labour supply with respect to the fare (or the frequency) is the weighted sum of the demand elasticities for car and public transport commuting, where the weights are the modal shares.

We realised the calibration exercise in R coding environment, and searched the targeted set of parameters using a Differential Evolution optimisation package (Mullen et al., 2011). The calibration was successful in the sense that both initial demand levels \( x_r \) and \( x_c \) as well as the updated demand values after price and frequency changes are recovered with high precision.

5 Optimal transport supply: Numerical results

In this section, we discuss numerical results for the three scenarios mentioned above. In each case we maximize the relevant social welfare function and determine the optimal policy variables. Optimisation is performed using a sequential quadratic programming algorithm. We make sure that the algorithm leads to global optima by (i) visually observing the contour plot of social welfare in the meaningful range of the policy variables, (ii) using a Monte Carlo approach of randomly selected starting values and comparing the optima found after local searches.

5.1 Scenario A: Isolated public transport service

This scenario considers optimal fares and frequency for public transport, assuming there is no competition with road transport at all\(^{18}\). Numerical results are provided in Figure 1. The

\(^{17}\)Note in Table 1 that vehicle size parameter \( s \) is calibrated such that under the baseline demand conditions and frequency \( f \), the resulting crowding density should be \( \phi = 3 \) passengers per square metre. Similarly, road capacity \( K \) is set such that car travel time \( t_c \) in the initial equilibrium should be two times the free-flow travel time \( t_f \). Road use related parameters are not used in Scenario A, while road toll \( \tau \) is set to zero in scenario B and endogenously optimised in scenario C.

\(^{18}\)The purpose of this scenario is partly pedagogical, as it illustrates the basic properties of the model without any influence of the competing car mode. It also allows comparison with the earlier literature of mono-modal public transport models. Finally, it serves as a crude approximation for very dense cities (e.g. Hong Kong, Tokio) where car commuting to the downtown area is hardly a reasonable alternative for high capacity public transport.
Table 1: Baseline simulation parameters

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>50,000</td>
<td></td>
<td>Number of potential (homogeneous) users</td>
</tr>
<tr>
<td>( L )</td>
<td>8</td>
<td>hour/day</td>
<td>Daily working hours</td>
</tr>
<tr>
<td>( \bar{L} )</td>
<td>24</td>
<td>hour/day</td>
<td>Length of workday</td>
</tr>
<tr>
<td>( h )</td>
<td>1.5</td>
<td>hour</td>
<td>Length of am/pm peak</td>
</tr>
<tr>
<td>( w )</td>
<td>100</td>
<td>$/day</td>
<td>Daily wage</td>
</tr>
<tr>
<td>( \eta )</td>
<td>1.04</td>
<td></td>
<td>Agglomeration elasticity</td>
</tr>
<tr>
<td>MCPF</td>
<td>1.066</td>
<td></td>
<td>Multiplier of marginal cost of public funds</td>
</tr>
</tbody>
</table>

Initial public transport supply

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>10</td>
<td>trains/hour</td>
<td>Rail frequency</td>
</tr>
<tr>
<td>( p )</td>
<td>4</td>
<td>$/trip</td>
<td>Rail fare</td>
</tr>
<tr>
<td>( \phi )</td>
<td>3</td>
<td>pass/m²</td>
<td>Rail crowding density ( \rightarrow ) veh. size(^{17})( (s) )</td>
</tr>
<tr>
<td>( t_v )</td>
<td>0.5</td>
<td>hour</td>
<td>In-vehicle rail travel time</td>
</tr>
<tr>
<td>( \delta_w )</td>
<td>1.2</td>
<td></td>
<td>Value of time multiplier for waiting time</td>
</tr>
<tr>
<td>( \delta )</td>
<td>0.15</td>
<td></td>
<td>Slope of crowding multiplier</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.6</td>
<td></td>
<td>Cost recovery ratio</td>
</tr>
<tr>
<td>( \omega )</td>
<td>0.6</td>
<td></td>
<td>Operational cost elasticity w.r.t. frequency</td>
</tr>
</tbody>
</table>

Initial road use parameters – Scenarios B and C

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>8</td>
<td>$/trip</td>
<td>Monetary operational cost of car use</td>
</tr>
<tr>
<td>( \tau )</td>
<td>2</td>
<td></td>
<td>Road (or fuel) tax</td>
</tr>
<tr>
<td>( t_f )</td>
<td>0.25</td>
<td>hour</td>
<td>Free-flow travel time by car</td>
</tr>
<tr>
<td>( t_c )</td>
<td>0.75</td>
<td>hour</td>
<td>Congested travel time ( \rightarrow ) road capacity(^{17})( (K) )</td>
</tr>
</tbody>
</table>

Initial demand

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_r )</td>
<td>0.2</td>
<td>roundtrip/day</td>
<td>Rail commuting demand / labour supply</td>
</tr>
<tr>
<td>( x_c )</td>
<td>0.2</td>
<td>roundtrip/day</td>
<td>Car commuting demand / labour supply</td>
</tr>
<tr>
<td>( \varepsilon_r )</td>
<td>-0.4</td>
<td></td>
<td>Price elasticity of rail demand</td>
</tr>
<tr>
<td>( \varepsilon_{rf} )</td>
<td>0.6</td>
<td></td>
<td>Frequency elasticity of rail demand</td>
</tr>
<tr>
<td>( \varepsilon_{cr} )</td>
<td>0.15</td>
<td></td>
<td>Cross-price elasticity of car use w.r.t. rail fare</td>
</tr>
<tr>
<td>( \varepsilon_{cf} )</td>
<td>-0.2</td>
<td></td>
<td>Cross-frequency elasticity of suburban car use</td>
</tr>
</tbody>
</table>
Consider the upper panel of Figure 1. The welfare maximising frequency is approximately 19 trains per hour, which implies that frequency was undersupplied in the initial equilibrium to which we calibrated the model (there were 10 trains per hour in the baseline model). The optimal fare of around $0.6 is well below its initial value as well (this was $4, see Table 1). The low fare is the combined result of the negative crowding externality, the positive agglomeration benefit, and the revenue generating mark-up. If the optimal fare was mainly driven by the marginal external crowding cost, the crowding density of 3 passengers per square metre in the initial equilibrium would imply significantly higher optimal fares. In fact, we found that the socially optimal crowding density is actually slightly higher than the reference value. The result suggests therefore that the positive agglomeration externality may neutralise a considerable part of the negative crowding externality as well as the mark-up resulting from costly public funds.

In the lower panel of Figure 1 we plot the paths of the optimal pair of fares and frequencies when varying four key parameters: (i) the operational cost elasticity, (ii) the agglomeration elasticity, (iii) the slope of the crowding multiplier function, and (iv) the tax revenue premium derived from the marginal cost of public funds.

Let us begin with the operational cost parameter $\omega$. Not surprisingly, Figure 1 shows that increasing the operational cost elasticity reduces the optimal frequency, but it also raises the fare. The former effect is stronger than the effect on the fare. In the range of $\omega \in (0.5, 1)$ the optimal fare increases by less than $2.

The crowding parameter $\delta$ has important effects for both frequency and fares. We vary $\delta$ between zero (i.e., no crowding inconvenience) and 0.35, which is the upper bound of the highest values in the empirical literature (Wardman and Whelan, 2011). Note that the optimal fare even drops below zero when crowding is not very important because, in that case, the agglomeration benefit dominates the crowding effect. When crowding is sufficiently important fares are positive; observe, however, that even for the highest parameter value considered the fare is less than the reference value of $4.

---

$^{19}$We also tested sensitivity with respect to the value of waiting time parameter, $\delta_w$. We found that the optimal frequency $f^*$ increases, while the fare $p^*$ decreases with $\delta_w$. However, the sensitivity is relatively limited: in the range of $\delta_w \in (1, 1.4)$, $f^*$ varies between 18.8 and 19.4 trains per hour, and $p^*$ remains between $0.66 and $0.56.

$^{20}$Technically, we sequentially recalibrate the operational cost function to make sure that the initial cost recovery ratio of $\rho = 0.6$ remains constant at various levels of the elasticity $\omega$. In practice this also means the recalibration of $z_0$ using equation (20).
Figure 1: Scenario A – Sensitivity of optimal supply for an isolated public transport service, with no modal substitution.
Interestingly, the relation between optimal frequency and the crowding parameter has an inverse U-shape. This result can be explained by changes in demand. As passengers’ sensitivity to crowding (δ) increases, the operator optimally reacts with higher frequency. This reduces the crowding externality for a constant number of passengers, but it also increases demand. For relatively low values of the crowding parameter, the former effect dominates, implying higher optimal frequency. At some high value, however, the negative impact of crowding inconvenience on commuting demand becomes so strong that it is not worth any more providing higher frequency for a dwindling mass of public transport users. The Mohring effect, i.e., density economies in waiting time, can be a secondary force playing a role in this phenomenon (see Bar-Yosef et al., 2013).

Now turn to the sensitivity with respect to the agglomeration elasticity. Earlier literature (reviewed by Melo et al., 2013) found that the majority of productivity elasticities with respect to agglomeration range between 1 and 1.08. Unsurprisingly, Figure 1 shows that the degree to which external agglomeration benefits can neutralise crowding depends heavily on this elasticity η. In the absence of agglomeration benefits (and given the baseline crowding parameter), the optimal fare equals $3. With agglomeration elasticities in the higher range of plausible values, normally found for high-tech and finance industries, the fare can drop below zero once again, fully neutralising the need for the internalisation of crowding externalities.

This is not the first paper that finds that negative transport taxes cannot be ruled out under agglomeration economies: Arnott (2007) states that “if tolling is the only instrument available to mitigate the distortions associated with congestion and agglomeration externalities, then, within the range of empirical estimates [of the agglomeration elasticity and congestion parameters], and for a particular model, the optimal toll might not only be substantially lower than the Pigouvian toll, but might even be negative, entailing a subsidy to urban travel”. Similarly, Brinkman (2016) infers from policy simulations with a spatial general equilibrium model of urban structure that “congestion pricing may have ambiguous consequences for economic welfare”, due to the offsetting externalities at work.

The agglomeration elasticity has a crucial impact on optimal frequency as well. In the range introduced above, the frequency is about 11 trains per hour (a headway of 5-6 minutes) if there is no productivity effect, and it increases to 30 trains per hour when η = 1.08. This suggests that considering agglomeration economies may be an important input for timetable setting; so far, this has not been observed in the literature.

Finally, we simulate the sensitivity of optimal fares and frequency with respect to the tax revenue premium. As intuition suggests, a higher tax revenue premium implies higher fares and lower frequencies, thus reducing the financial deficit of service provision. The magnitude of changes in the range of tax premia considered (between 0 and 30%) is substantial; depending on this parameter, the optimal fare may vary between -$0.85 and $2.6.

Figure C.1 in Appendix C provides a number of additional results. It graphically illustrates the impact of the fare and the frequency on public transport demand, on crowding, on wages, on the cost recovery ratio of public transport, and on the government’s deficit (expressed as the ‘tax’ per person needed to cover the deficit). For example, panel (a) in Figure C.1 plots individual daily rail demand x_r in function of the rail fare p and frequency
Each contour line in the figure represents combinations of the fare and the frequency that imply the same demand level (a similar interpretation holds for the other figures). We also indicate the fare and the frequency at the initial equilibrium ($p = 4$, $f = 10$) and at the social optimum ($p = 0.6$, $f = 19$) on the figure. We learn from this plot that the initial rail demand of $x_r = 0.2$ increases to just above 0.35 when we move to the social optimum.

Panel (b) of Figure C.1 shows that crowding does not change significantly; in both equilibria density is close to 3 passengers/m$^2$. This implies that the marginal external crowding cost is relatively stable. It is further remarkable that, despite the observed sensitivity of optimal supply with respect to agglomeration, the equilibrium wage moves in the expected direction, but the change is relatively small. Although agglomeration forces are highly relevant for setting fares and frequencies, the effect on individual wages is not dramatic. Finally, the optimal combination of fares and frequencies has a direct impact on the level of subsidisation as well. As part of the simulation we divide the financial loss (or surplus) of public transport provision by $N$ to get the mean lump-sum tax (or subsidy) that households would have to pay (or receive) to balance the operator’s budget. Panel (d) shows that moving from the initial equilibrium to the social optimum implies a considerable increase in per capita tax from $1 to around $3.5. This reduces the cost recovery ratio of public transport from 0.6 to less than 0.1.

5.2 Scenario B: Modal competition with untolled road use

In the second scenario we turn to a two-mode setting in which car use and rail use are treated as imperfect substitutes\textsuperscript{21}. We consider optimal public transport policies (fares and frequencies) conditional on keeping the road tax constant at its reference value of $\tau = $2.

Results are summarised in Figure 2 which has the same interpretation as Figure 1 above. First consider the upper panel. This suggests that optimal fares are lower than in the previous case, amounting to about $0.25 per trip. The optimal second-best frequency is slightly lower than in the previous scenario; it amounts to about 18 trains per hour. These second-best results are partly driven by the inefficiency of under-priced road congestion, and the fact that productivity benefits are generated in both modes.

Considering the scope of our analysis, the most remarkable feature of this scenario is that the impact of the agglomeration elasticity on the optimal fare and frequency is much smaller in the presence of a sub-optimally priced alternative. The variation in $p$ in the interval $\eta \in \{1, 1.08\}$ for the agglomeration elasticity is limited to just about $0.2. This outcome can be explained both technically and intuitively. Reconsider the theoretical expression of the optimal fare in equation (10). As a fare increase has the opposite effect on demand for car and rail commuting, i.e. $\partial x_c / \partial p > 0$ while $\partial x_r / \partial p < 0$, it is impossible to internalise the agglomeration productivity benefits of rail commuting without losing some of the external benefits that car commuters generate. This is due to the assumption that agglomeration is not mode-specific externality: it is the total number of workers (commuters) that determines

\textsuperscript{21}This follows from the specification of the model in (15)–(18). With our baseline demand elasticities, $\frac{\partial x_r}{\partial p} / \frac{\partial x_c}{\partial p} = \frac{\epsilon_{rx}}{\epsilon_{rx}} = -0.375$. That is, for each rail user diverted by a marginal change in the rail fare, substitution generates 0.375 additional car trips.
Figure 2: Scenario B – Sensitivity of optimal public transport supply with untolled road use.
agglomeration effects, not their modal distribution. Therefore, inducing a mode shift has conflicting effects on agglomeration in the two modes, and only the net effect matters. The extent to which these opposing effects cancel each other out depends on the degree of substitution between the two modes, i.e., on the own and cross-price elasticities in the model. Our model features imperfect substitution, and is calibrated to the moderate cross-price elasticity of $\varepsilon_{cr} = 0.15$. Still, our results suggest that the majority of the productivity benefits of induced rail demand is neutralised by lost car commuting, and therefore the strength of agglomeration effects does not have a large impact on pricing. The same holds for the agglomeration effect on frequency: optimal frequencies range between 16.5 and 19.5 vehicles per hour, a much smaller range than observed in Figure 1.

Interestingly, the sensitivity of the optimal fare with respect to the crowding cost parameter indicates that $p^*$ actually slightly decreases when sensitivity to crowding becomes more important (higher $\delta$). Remember that this relationship was positive in scenario A, and theory also suggests that the fare should increase with $\delta$, as it is one of the key determinants of the marginal external crowding cost of a trip. However, the detailed simulation results reveal that the low fare is driven by another determinant of the external crowding cost: the equilibrium demand for rail travel is low when passengers are sensitive to crowding. This, combined with the high frequency provided by the rail operator lead to a crowding density as low as 1.95 pass/m$^2$ at $\delta = 0.3$. At the same time, congestion on the parallel road is high, and the cost recovery ratio of the rail service is close to zero. By contrast, when sensitivity to crowding is negligible, the operator does not need to provide high frequency and low fares to divert drivers off the road, and road congestion is eliminated to a much greater extent. In this case, crowding in public transport is unreasonably high (passengers are insensitive to it), and the corresponding cost recovery ratio is way above zero in optimum. In summary, the crowding parameter is an important determinant of substitution between the two modes.

Similarly, the impact of varying scale economies in operational costs for the optimal fare may seem counterintuitive at first sight. One would expect that density economies should lead to high optimal frequency, therefore lower crowding, and consequently lower fares. This is what happened in scenario A. However, what we see in Figure 2 is the opposite. The intuition is that cheap frequency and low crowding make public transport attractive for car drivers, even without ‘discounts’ on the pricing side. By contrast, if high frequency is expensive due to the absence of density economies, then crowding is inevitable, and reducing fares is the only way to alleviate road congestion. This argument is supported by the detailed simulation results: the density of crowding is $\phi = 2.41$ pass/m$^2$ at $\omega = 0.5$ (strong density economies in operational costs), and 4.07 pass/m$^2$ when $\omega = 0.95$. As a consequence, in the latter case the operator achieves a significantly higher cost recovery ratio despite the absence of scale economies. Note that these seemingly counter-intuitive results are due to the availability of an under-priced substitute.

In Figure 2 we add another set of optima representing various levels of road capacity. Restricting road capacity below the reference level (which was $K = 3500$) leads to higher fares combined with higher rail frequencies. It is easy to explain the reason behind high frequency in this case: the road section is heavily congested, and therefore the need for the
public transport service to induce a modal shift. More interestingly, the optimal crowding density is almost independent of the available road capacity, because the operator combines high frequency with high fares. A potential explanation is that heavy road congestion makes $x_c$ less price elastic, and therefore under-priced public transport can no longer be used as a second-best tool to alleviate road congestion.

Of course, the optimal fare and frequency levels we report in Figure 2 depend heavily on the road transport tax we selected. We do not show this explicitly in the figure, but we found that increasing the exogenous road toll shifts the sensitivity curves in the lower panel of Figure 2 upwards, reflecting higher public transport fares; however, their overall shape does not change significantly. Figure C.2 provides more simulation results in a visual form, including the share of public transport use in function of fares and frequencies.

Finally note that, although we focused on the role of agglomeration benefits when car use is under-priced, a slight reinterpretation of our findings also yields an important policy message for road pricing. Arnott (2007) finds that the optimal congestion charge should diverge from the Pigouvian level, to consider the agglomeration externality. Our analysis suggests this is not the case under strong modal substitution, unless agglomeration is considered in public transport pricing as well. In other words, there is no reason to reduce congestion charges due to agglomeration, as long as public transport as a close substitute is not correctly priced.

5.3 Scenario C: Modal competition with optimal road pricing

In the third scenario we stick to the assumption that car and rail use are imperfect substitutes, but now jointly optimise fares, road tolls and frequency of public transport. The three-dimensional optimisation implies that it is no longer feasible to summarise the results on two-dimensional contour plots. We therefore summarise the results in Table 2.

The most important result from the viewpoint of the present paper is that the agglomeration externality is once again a key determinant of the optimal public transport fare. As opposed to scenario B, fares are quite sensitive to variations in the agglomeration elasticity. This sensitivity is comparable to that of varying the tax revenue premium. With extremely strong agglomeration benefits and costless tax collection, negative fares cannot be excluded. Agglomeration economies are highly relevant in this scenario because the productivity externality of commuting is now internalised in the two modes separately. Thus, the road demand related congestion and agglomeration components disappear from the fare formula in equation (10), leading to $p^*_{r,t}$ in (13).

This setting shows similarities with the isolated public transport model in other respects as well. The optimal fare is an increasing function of the crowding multiplier, due to its dependence on the marginal external crowding cost. Density economies in operational costs imply lower fares due to the decreasing average cost of serving higher demand. Finally, $p^*$ is just weakly dependent on the available road capacity, because optimal congestion pricing eliminates the second-best role of the public transport fare. Nevertheless, the optimal frequency does react to variations in $K$: the simulation suggests that if road capacity is limited, then the operator supplies higher frequency to accommodate the increased residual demand for rail commuting.
### Table 2: Sensitivity of optimal transport supply with respect to key parameters.

<table>
<thead>
<tr>
<th>f</th>
<th>p</th>
<th>τ</th>
<th>η</th>
<th>δ</th>
<th>ω</th>
<th>(β/κ) − 1</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>fare</td>
<td>toll</td>
<td>agglomeration</td>
<td>crowding</td>
<td>op. cost</td>
<td>tax premium</td>
<td>road cap.</td>
</tr>
<tr>
<td>Optimum with initial parameters</td>
<td>16.40</td>
<td>2.49</td>
<td>4.11</td>
<td>1.04</td>
<td>0.15</td>
<td>0.6</td>
<td>0.066</td>
</tr>
<tr>
<td>Agglomeration elasticity</td>
<td>13.55</td>
<td>4.91</td>
<td>6.38</td>
<td>1.00</td>
<td>14.91</td>
<td>3.74</td>
<td>5.27</td>
</tr>
<tr>
<td>Crowding multiplier</td>
<td>12.06</td>
<td>1.38</td>
<td>2.62</td>
<td>0.05</td>
<td>14.64</td>
<td>2.03</td>
<td>3.47</td>
</tr>
<tr>
<td>Operational cost elasticity</td>
<td>21.25</td>
<td>1.86</td>
<td>3.20</td>
<td>0.50</td>
<td>18.53</td>
<td>2.19</td>
<td>3.67</td>
</tr>
<tr>
<td>Tax revenue premium</td>
<td>19.97</td>
<td>-2.00</td>
<td>0.07</td>
<td>0</td>
<td>18.32</td>
<td>0.07</td>
<td>1.94</td>
</tr>
<tr>
<td>Road capacity</td>
<td>18.81</td>
<td>2.20</td>
<td>3.73</td>
<td>3000</td>
<td>17.76</td>
<td>2.33</td>
<td>3.89</td>
</tr>
</tbody>
</table>

### 5.4 Agglomeration, fares and frequencies: concluding remarks

Three final remarks conclude this section. First, the comparison to the three scenarios A, B and C analyzed in detail in this section leads to the policy message that agglomeration agglomeration elasticities become relevant for public transport pricing if (i) the externalities associated with car use are appropriately priced, or (ii) the degree of substitution between the two modes is negligible either due to generic user preferences or the scarcity of road space. This finding is driven by the reasonable assumption that agglomeration externalities are not mode specific. Therefore, if road tolls do not internalise agglomeration benefits, capturing them via low public transport fares is inappropriate. This would generate a modal shift from car to public transport, so that the agglomeration benefits of more rail commuting will be largely neutralised by the loss of agglomeration benefits due to reduced commuting by car.

Second, note that all three numerical scenarios are calibrated with equal modal shares.
in the initial equilibrium. As part of our sensitivity analysis, we re-calibrated the model with a range of unequal initial modal shares, keeping total demand (i.e., $x_r + x_c$) and other parameters constant. This experiment suggests that cities with low initial public transport patronage have lower fares and lower frequencies in the welfare optimum, compared to cities where public transport has a large modal share in the baseline equilibrium. A word of caution must be expressed here, however. In these exercises we kept all other parameters constant, while it is very unlikely that cities with substantially different modal split do not differ in terms of demand elasticities and crowding and congestion levels, for example.

Finally, we ignored heterogeneity in commuters’ contribution to agglomeration. For example, suppose there are multiple types of users, and let their contribution to productivity differ no matter which mode they travel with. Unequal productivity leads to unequal wages, and modal shares within each group might differ, so that at the aggregate level the average car and the average transit user may have a different contribution to agglomeration. What one expects to happen within the framework of the model is that the “productivity” component in the optimal pricing formula becomes greater in magnitude for the mode that is predominantly used by more productive commuters. For example, if more productive workers are public transport users, then this sort of heterogeneity would lead to lower fares compared to the case of equal modal splits in each group. Still another type of heterogeneity arises when productivity is coupled with some degree of spatial segregation. If modal shifts go along with changes in residential and working locations then pricing reforms can indeed lead to a densification of the city. Unfortunately, due to the absence of a land market in our model, we cannot predict the direction of changes this would imply for our results.

6 Agglomeration and cost recovery in public transport

In this final section we zoom in on the implications of agglomeration for cost recovery in public transport. Figure 3 depicts the cost recovery ratio in public transport for the three scenarios A, B and C. The revenues collected from road taxes are excluded, as public transport subsidies are normally considered independent of the road sector in policy debates. Figure 3 shows that the agglomeration elasticity has a negative impact on the optimal cost recovery ratio in all scenarios. The negative effect is especially strong in scenarios A and C, i.e., when modal substitution is absent or road use is efficiently priced. Within the range of agglomeration elasticities we consider in this paper, the degree of self-financing may drop by not less than 80% in these cases (in scenario C it drops from 84% to just below zero). Despite the nonlinear specification of most model components, the shape of the optimal cost recovery reflects an almost linear relationship with $\eta$, which may be useful in practical policy analysis. The rule of thumb that can be derived for an isolated public transport service, for example, is that a percentage point increase in the agglomeration elasticity leads to around 12% lower cost recovery ratio.

We also see, however, that the self-financing ratio is almost independent of agglomeration in the case of fixed and lower-than-optimal road taxation. Under the current set of parameters ($\tau = $2) cost recovery is low, so that a substantial share of operational costs of public
transport provision should be covered by public subsidies. The sensitivity analyses show that higher (lower) values of \( \tau \) shift the \( \rho(\eta) \) curve of scenario B upwards (downwards), but the slope of the curve does not change significantly. In the extreme case of completely untaxed road use, second-best public transport fares and the corresponding cost recovery ratio may even become negative.

Figure 3: Cost recovery ratio and its dependency on agglomeration economies and substitution with congestible car use.

Table 3: Cost recovery ratio (\( \rho \)) and other performance metrics with separate and joint optimisation of tolls, fares and frequencies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( f )</th>
<th>( p )</th>
<th>( \tau )</th>
<th>( x_r )</th>
<th>( x_c )</th>
<th>( \phi )</th>
<th>( t_c/t_f )</th>
<th>( w )</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units</td>
<td>1/h</td>
<td>$</td>
<td>$</td>
<td>trip/day</td>
<td>trip/day</td>
<td>pax/m²</td>
<td>-</td>
<td>$/day</td>
<td>%</td>
</tr>
<tr>
<td>Initial equilibrium</td>
<td>10.00</td>
<td>4.00</td>
<td>2.00</td>
<td>0.20</td>
<td>0.20</td>
<td>3.01</td>
<td>2.04</td>
<td>100.02</td>
<td>0.602</td>
</tr>
<tr>
<td>Optimal ( \tau )</td>
<td>10.00</td>
<td>4.00</td>
<td>6.15*</td>
<td>0.26</td>
<td>0.14</td>
<td>3.87</td>
<td>1.27</td>
<td>100.02</td>
<td>0.773</td>
</tr>
<tr>
<td>Optimal ( f, p )</td>
<td>17.94*</td>
<td>0.27*</td>
<td>2.00</td>
<td>0.34</td>
<td>0.14</td>
<td>2.86</td>
<td>1.21</td>
<td>100.71</td>
<td>0.048</td>
</tr>
<tr>
<td>Optimal ( f, p ), Pigouvian ( \tau )</td>
<td>17.41*</td>
<td>0.54*</td>
<td>2.08*</td>
<td>0.33</td>
<td>0.14</td>
<td>2.87</td>
<td>1.25</td>
<td>100.68</td>
<td>0.097</td>
</tr>
<tr>
<td>Optimal ( f, p, \tau )</td>
<td>16.40*</td>
<td>2.49*</td>
<td>4.11*</td>
<td>0.32</td>
<td>0.13</td>
<td>2.89</td>
<td>1.19</td>
<td>100.46</td>
<td>0.440</td>
</tr>
</tbody>
</table>

* Sensitivity w.r.t. agglomeration elasticity

\( \eta = 1 \) | 13.55* | 4.91* | 6.38* | 0.28 | 0.13 | 3.05 | 1.20 | 100.00 | 0.846 |
| \( \eta = 1.08 \) | 19.84* | -0.49* | 1.37* | 0.36 | 0.13 | 2.73 | 1.19 | 101.69 | -0.088 |

*: optimised decision variables
Finally, consider Table 3. It provides an overview of numerical results for rail and car demand, crowding, wages, and cost recovery ratios for public transport. All results are based on the reference set of model parameters (operational cost elasticity, agglomeration elasticity, crowding parameter, cost of fund parameter). Beside the reference equilibrium, which serves as benchmark, and scenarios B and C, we also report what happens when the toll is optimised for given reference values of public transport fares and frequency.

In the reference equilibrium, rail frequency was undersupplied while the fare was higher than socially optimal; road use was under-priced. If only the road tax is optimised with constant public transport supply, this reduces the intensity of car commuting by 30% and considerably increases the density of crowding in public transport. As fares are fixed, the mode shift to public transport boosts the cost recovery ratio of rail operations, measured as the amount of fare revenues relative to operational costs, up to 77%. If only public transport decision variables \( p \) and \( f \) can be optimised with fixed \( \tau = $2 \), this leads to significantly higher frequency and very low fares; the degree of self-financing drops to just around 5%. In this second-best optimum road congestion relief is actually stronger than with optimal road pricing, and the increase in aggregate labour supply raises the average wage for all commuters.

In Table 3 we also consider another potentially relevant policy scenario, viz., the case of a naïve road toll set to internalise the congestion externality only, neglecting agglomeration effects\(^{22}\). Equation (11) in the theoretical analysis predicts that this Pigouvian congestion charge may be higher as well as lower than its first-best optimum, depending on the relative magnitude of the agglomeration externality and the mark-up due to costly public funds. With the present parameters, the Pigouvian toll turns out to be slightly higher than the fixed toll in scenario B, but lower than the first best \( \tau^* \) in scenario C. The optimal public transport supply varies in line with prior expectations. Finally, if all three decision variables are simultaneously optimised, the cost recovery ratio returns to an intermediate level of 44%. The intensity of car use is the lowest in this case, but rail commuting is also limited compared to the previous case with fixed road taxes.

In the last two rows of Table 3 we zoom in on the sensitivity of the results with respect to the agglomeration elasticity when all three decision variables are selected optimally; thus, this sensitivity analysis is to be compared with the final row above the horizontal line in Table 3. We noticed before that fares and frequency are quite sensitive to agglomeration. This generates a huge sensitivity of the optimal self-financing ratio as well. Its values for the extremes of \( \eta \) in the interval \((1, 1.08)\) are 84% and -8% in the last column of the table, respectively. One interpretation is that high agglomeration benefits ‘justify’ low cost recovery ratios in public transport. By contrast, crowding in public transport \( (\phi) \), the degree of road congestion \((t_c/t_f\) is the equilibrium travel time relative to the free-flow time\), and wages are much less sensitive to agglomeration under optimal transport supply.

\(^{22}\)We thank one of the referees for suggesting this scenario.
7 Conclusions

The purpose of this paper was to investigate the impact of agglomeration externalities on socially optimal public transport fares and frequency. We develop a simple model in which commuters have the choice between car and public transport use. The former is subject to road congestion, the latter potentially suffers from crowding. Overall labour supply generates agglomeration externalities. Using a fairly standard representation of public transport operations, we studied the role of agglomeration benefits for socially optimal transport prices and public transport frequency and, by implication, the effect of agglomeration on the cost recovery ratio of the public transport system. A numerical application illustrated the theoretical findings.

We find that the effect of agglomeration benefits on fare and frequency setting in public transport crucially depends on the pricing of car use. When the price of car use reflects its net social cost including the congestion externality, then agglomeration economies have a large effect on the optimal fare-frequency combination and, by implication, on the optimal cost recovery ratio in public transport. One interpretation is that high agglomeration benefits ‘justify’ low cost recovery ratios in public transport, even with optimally priced car use (hence, absent second-best considerations related to suboptimal pricing of congestion). A numerical version of the theoretical model implies very low cost recovery ratios when agglomeration elasticities are near the upper bound of estimates reported in the empirical literature. For the most usual estimate of the agglomeration elasticity, $\eta = 1.04$, the cost recovery ratio we find is 44%.

However, when road use is under-priced because congestion externalities remain untaxed, the sensitivity of fares and frequencies to agglomeration benefits largely disappear. The reason is that agglomeration externalities are not mode-specific, so that, if the productivity benefit cannot be internalised on the road, then any mode shift induced by lower public transport fares has opposing agglomeration effects on the two modes. This finding of the paper has important policy implications for road pricing as well. Arnott (2007) finds that the optimal congestion charge should diverge from the Pigouvian level, to consider the agglomeration externality. Our analysis suggests this is not the case under strong modal substitution, unless agglomeration benefits are considered in public transport pricing as well.

This paper opens up a wide range of future research topics. First, we considered commuting to work as the only trip purpose, while public transport networks serve a number of additional trip purposes in the urban economy. Here agglomeration benefits may arise as well (Rosenthal and Strange, 2004), for example in consumption enabled by shopping and leisure trips. Second, real-world public transport normally serves multiple spatio-temporally dif-

---

23 A realistic approach to take into account non-commuting trips – many, but not all, of which are off-peak trips – would suggest extending the model to two periods. Van Dender (2003) considers such a two-period model, distinguishing peak and off-peak trips by car and public transport, and capturing both commuting and non-commuting. Although his model ignores agglomeration effects, his results suggest higher public transport fares in the peak period than in the off-peak, where the optimal fare might well be zero. Introducing agglomeration is unlikely to strongly affect the relative fares in the two periods, although it might affect the absolute levels.
differentiated markets with the same capacity (Hörcher and Graham, 2018), as opposed to our stylised static model with only one origin-destination pair. Imbalances in demand and heterogeneity in trip purposes may have severe impact on supply variables and subsidisation. Third, users may be heterogeneous in other characteristics as well, including their contribution to productivity, income, and modal preferences (David and Foucart, 2014). Four, the model would yield additional insights if residential location (see e.g. Borck, 2019), city size and departure times (Fosgerau and Kim, 2019) could be endogenised. Assuming that car and public transport users on average differ in income and contribution to agglomeration, the productivity effect of inducing labour supply through transport policy might also differ. Thus, the determination of the efficient subsidy level in a particular urban area may be a more complex task than what this paper presents as a numerical example. However, even this simple example highlights that agglomeration is an indispensable factor of optimal urban public transport provision, and external agglomeration benefits have to be considered when subsidies for such services are determined.

Acknowledgement

An earlier version of this paper was presented at the 2018 International Transport Economics Association conference (Hong Kong, June 2018). We are also grateful to two referees and an editor of this journal for many constructive remarks that much improved the final version of this paper. We take responsibility for remaining shortcomings.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendices

A Derivation of optimal fares and frequencies

This section derives the optimal supply rules of equations (12) and (10) from the constrained welfare maximisation problem in equation (9). For the sake of tractability we repeat this Lagrangian here:

\[ \Lambda_s = N \cdot V(p, f, \tau) + \beta [N(x_r p + x_c \tau) - z(f)]. \]

Let us first consider that indirect utility function \( V \) is by definition the value of \( u \) at the utility maximising level of commuting by the two modes:

\[ V(p, f, \tau) = U(y^*, l^*) - x_r \cdot \varphi(Nx_r f^{-1}), \]
Thus, equation (A.2) becomes
\[
\frac{\partial V}{\partial f} = U_y \left[ (w - p + x_r w'(N x) N) \frac{\partial x_r}{\partial f} + (w - c - \tau + x_c w'(N x) N) \frac{\partial x_c}{\partial f} \right] \\
- U_t \left[ (L + t_r) \frac{\partial x_r}{\partial f} + x_r t_r'(f) + (L + t_c + x_c t_c'(N x_c) N) \frac{\partial x_c}{\partial f} \right] \\
- (\varphi + x_r \varphi_{x_r}) \frac{\partial x_r}{\partial f} - x_r \varphi_f,
\]
where \( \varphi_{x_r} > 0 \) and \( \varphi_f < 0 \), because crowding discomfort increases with the number of commuters and decreases with rail frequency. Note that marginal utilities can be replaced by \( \kappa \) and \( \mu \), the Lagrange multipliers of the consumer’s problem. Using our assumptions on the functional forms of congestion technology and crowding inconvenience\(^{24}\), equation (A.1) simplifies to
\[
\frac{\partial V}{\partial f} = \frac{\partial x_r}{\partial f} \left[ \kappa (w - p + x_r w' N) - \mu (L + t_r) - \varphi - x_r \varphi_{x_r} \right] \\
+ \frac{\partial x_c}{\partial f} \left[ \kappa (w - c - \tau + x_c w' N) - \mu (L + t_c + x_c t_c' N) \right] + \mu 0.5 f^{-2} x_r - x_r \varphi_f.
\]

Now, this expression can be further simplified by neutralising some of its components with the first order conditions of the utility maximisation problem in equation (4) with respect to \( x_r \) and \( x_c \):
\[
\kappa (w - p) - \mu (L + t_r) - \varphi = 0, \\
\kappa (w - c - \tau) - \mu (L + t_c) = 0.
\]

Thus, equation (A.2) becomes
\[
\frac{\partial V}{\partial f} = \frac{\partial x_r}{\partial f} \left[ \kappa x_r w' N - x_r \varphi_{x_r} \right] + \frac{\partial x_c}{\partial f} \left[ \kappa x_c w' N - \mu x_c t_c' N \right] + \mu 0.5 f^{-2} x_r - x_r \varphi_f. \tag{A.4}
\]

Following the same logic, the impact of fare setting on indirect utility can be expressed first as
\[
\frac{\partial V}{\partial p} = U_y \left[ (w - p + x_r w'(N x) N) \frac{\partial x_r}{\partial p} + (w - c - \tau + x_c w'(N x) N) \frac{\partial x_c}{\partial p} - x_r \right] \\
- U_t \left[ (L + t_r) \frac{\partial x_r}{\partial p} + (L + t_c + x_c t_c'(N x_c) N) \frac{\partial x_c}{\partial p} \right] - (\varphi + x_r \varphi_{x_r}) \frac{\partial x_r}{\partial p}.
\]

The identical steps of algebraic manipulation lead to
\[
\frac{\partial V}{\partial p} = -\kappa x_r + \frac{\partial x_r}{\partial p} \left[ \kappa x_r w' N - x_r \varphi_{x_r} \right] + \frac{\partial x_c}{\partial p} \left[ \kappa x_c w' N - \mu x_c t_c' N \right], \tag{A.6}
\]
\(^{24}\)In particular, we use that \( t_v(N x_c) = \alpha N x_c \) leads to \( \partial t_v/\partial x_c = t_v'(N x_c) N = \alpha N \), while in case of rail travel times \( t_r(f) = t_v + 0.5 f^{-1} \) implies \( t_r'(f) = -0.5 f^{-2} \).
in which some of the expected externality components, i.e. agglomeration, crowding and congestion externalities, can be recognised immediately. Finally,

\[
\frac{\partial V}{\partial \tau} = -\kappa x_c + \frac{\partial x_r}{\partial \tau} \left[ \kappa x_r w'N - x_r \varphi x_r \right] + \frac{\partial x_c}{\partial \tau} \left[ \kappa x_c w'N - \mu x_c t'N \right].
\]  \hspace{1cm} (A.7)

We can proceed towards the optimal supply variables by taking first order conditions of the operator’s constrained optimisation problem.

\[
\frac{\partial \Lambda_s}{\partial f} = N \frac{\partial V}{\partial f} + \beta \left[ N p \frac{\partial x_r}{\partial f} + N \tau \frac{\partial x_c}{\partial f} - z'(f) \right] = 0
\]

\[
\frac{\partial \Lambda_s}{\partial p} = N \frac{\partial V}{\partial p} + \beta N \left[ x_r + p \frac{\partial x_r}{\partial p} + \tau \frac{\partial x_c}{\partial p} \right] = 0
\]

\[
\frac{\partial \Lambda_s}{\partial \tau} = N \frac{\partial V}{\partial \tau} + \beta N \left[ p \frac{\partial x_r}{\partial \tau} + x_c + \tau \frac{\partial x_c}{\partial \tau} \right] = 0
\]  \hspace{1cm} (A.8)

After plugging equations (A.4), (A.6) and (A.7) into these first order conditions, algebraic rearrangement leads to the optimal fare, toll and frequency formulae of equations (10), (11) and (12).
### B Commuting demand in the simulation framework

Let us consider the utility maximisation problem in the simulation framework, assuming that average generalised time $g_r$ and road travel time $t_c$ are exogenous to the consumer. The problem is:

$$
\max_{x_r, x_c} u = U((w - 2p)x_r + (w - 2c - 2\tau)x_c, \bar{L} - (L + 2t_r)x_r - (L + 2t_c)x_c) - \varphi(2x_r \cdot g_r, 2x_c \cdot t_c).
$$

(B.1)

The first-order condition of utility maximisation with respect to $x_r$ is

$$
\frac{\partial u}{\partial x_r} = U_y(w - 2p) - U_l(L + 2t_r) - \frac{\partial \varphi}{\partial T_r} 2g_r = 0,
$$

(B.2)

in which the derivatives of $U$ and $\varphi$, using the quadratic specifications defined by equations (15) and (16), are

$$
U_y = \alpha_y + \beta_y \cdot y + \gamma \cdot l,
$$

$$
U_l = \alpha_l + \beta_l \cdot l + \gamma \cdot y,
$$

$$
\frac{\partial \varphi}{\partial T_r} = \alpha_r + \beta_r \cdot T_r + \gamma_t \cdot T_c = \alpha_r + \beta_r 2x_r g_r + \gamma_t 2x_c t_c.
$$

(B.3)

The equivalent first-order condition for car commuting is

$$
\frac{\partial u}{\partial x_c} = U_y(w - 2c - 2\tau) - U_l(L + 2t_c) - \frac{\partial \varphi}{\partial T_c} 2t_c = 0,
$$

(B.4)

in which

$$
\frac{\partial \varphi}{\partial T_c} = \alpha_c + \beta_c \cdot T_c + \gamma_t \cdot T_r = \alpha_c + \beta_c 2x_c t_c + \gamma_t 2x_r g_r.
$$

(B.5)

Thus, (B.2) and (B.4) give a system of two unknowns ($x_r$ and $x_c$) in function of prices, the rail frequency, travel times and exogenous parameters. In the numerical implementation of the model we derive equilibrium demand levels by iteratively solving this system.
C Detailed simulation results

Figure C.1: Scenario A – Travel conditions and financial results in function of supply variables.
Figure C.2: Scenario B – Travel conditions and financial results in function of supply variables.
References


Gershenson, S. (2013), ‘The causal effect of commute time on labor supply: Evidence from a natural experiment


