The optimal speed of rail:
Velocity and its impact on fares, frequencies, and subsidies

Daniel Hörcher[1], Roger Vickerman[2], and Daniel J. Graham[3]

[1]Research Associate, Imperial College London, United Kingdom
[2]Professor, University of Kent and Imperial College London, United Kingdom
[3]Professor, Imperial College London, United Kingdom

Short summary

Should governments invest in high-speed rail? Do improvements in the conventional railway network deliver better value for money? This dilemma is among the leading themes of transport policy in many countries around the world, including the US, the UK, China, and several member states of the EU. In this paper we argue that the debate on railway policy should not be degraded into a binary decision between high-speed and conventional rail: we treat speed as a continuous decision variable in the optimisation of intercity transport provision, together with service frequency, pricing, and the level of public funding. Among several other policy-relevant results, we derive an analytical optimality rule for the welfare maximising speed and explore its interplay with service frequency and marginal cost pricing in a series of numerical simulations. Under a reasonable set of parameters, the model reveals previously unseen characteristics of optimal public transport provision: when speed is endogenous, the optimal frequency decreases with demand, and high-speed services are associated with higher fares to ensure the efficient use of resources.

Keywords: high-speed rail; pricing and capacity optimization; subsidies; transport economic and policy

1 Introduction

The commercial speed of passenger rail services varies on a wide range between trams travelling at the speed of a cyclist in dense urban areas and high-speed trains that may reach 360 km/h or more in daily operations. This is a unique feature of railways. The speed of road-based public transport is limited by traffic conditions and strict safety regulations that are standard around the world, and commercial aircraft's cruising speed is also bounded into a narrow interval by aerodynamic efficiency. By contrast, railways do not share the right of way with other modes, and technology enables safe and efficient operations at any speed within the range indicated above, especially if interactions with the built and natural environment are mitigated with dedicated infrastructure.

Policy debates about the speed of new railway infrastructure are intense in many countries. Such debates often simplify into a binary choice between loosely defined *high-speed* and *conventional* rail technologies.[1] Recent examples include HS2 in the United Kingdom (Vickerman, 2018), the rapid expansion of the high-speed rail network in China (Liu et al., 2023), and several mature plans to launch high-speed rail in the US (Button, 2012), including Florida and California.

Speed is the basis of a significant part of the user benefits in terms of the value of time savings. There has been a continuing attempt to increase the maximum speeds on high-speed rail lines as a means of creating greater user benefits albeit at the expense of higher costs. However, running speed is not the only determinant of service quality, and little is known about the interplay between velocity and other economic characteristics of rail service supply. The backbone of the literature of mass transit economics considers service frequency (Mohring, 1972; Chang & Schonfeld, 1991) and vehicle size (Jansson, 1980; Jara-Díaz & Gschwender, 2003) as key variables representing technological decisions in service provision, and investigates their impact on optimal pricing and

[1]High-speed and conventional rail are sometimes distinguished by the threshold speed of 200 km/h in engineering practice. The terminology *very high-speed rail* is sometimes used for services above 250 km/h.

subsidies (Börjesson et al., 2017; Hörcher & Graham, 2018). Travel times are sometimes endogenous with respect to stop spacing and demand, to capture the duration of boarding and alighting (Sun et al., 2014). Mixed traffic models with a congestion interaction between cars and public transport are also regular in the literature (Tirachini et al., 2014; Hörcher & Tirachini, 2021).

The speed of movements between stops is rarely treated as a standalone decision variable in existing public transport studies. To the best of our knowledge, the only exception is a paper by Tirachini & Hensher (2011) in which they consider running speed as an explicit policy outcome linked to the level of infrastructure investment in urban bus transport. As part of an extensive numerical simulation model, they find that speed is an increasing concave function of demand within the range of 30 to 70km/h. The more pronounced role of velocity in rail service provision motivates us to investigate the economics of speed more systematically by (i) deriving the analytical rule of the welfare-maximising speed and (ii) uncovering the link between commercial speed, service frequency, vehicle size, and the welfare economic and financial performance of public transport provision.

Speed has a crucial impact on several components of standard public transport models. Naturally, in-vehicle travel time savings can be realised by increasing the train speed. Also, due to the endogeneity of cycle times, the fleet size required to provide a given service frequency decreases with speed. On the other hand, the unit cost of capacity increases with commercial speed due to technological complexity, energy consumption, and externalities such as noise. These intuitive insights suggest that the *socially optimal speed* is a non-trivial function of (i) demand conditions and user preferences, (ii) technological parameters, (iii) the financial constraint imposed on the operator, and (iv) substitution with and pre-existing distortions in competing modes, e.g. aviation.

In this research we construct a microeconomic model of intercity transport supply in which the speed of service provision appears as an explicit supply-side variable.

2 Methodology

Let us consider a public transport (e.g. railway) service between an origin and a destination station, with no intermediate stops. The service covers a travel distance d within t hours, such that the average velocity is defined as $v = d/t$. In this model we neglect vehicle dynamics including the acceleration and deceleration of trains,[2] and assume that v, which is an average speed in reality, is the only velocity-related decision variable in the design problem. The cycle time of this service is

$$\sigma(v) = 2t(v) + t_0 = \frac{2d}{v} + t_0, \tag{1}$$

where t_0 is an exogenous dwell time at the two terminals, capturing the time required for boarding and alighting and any technological processes before the train departs in the opposite direction. We express service frequency f as the number of train departures per hour, so that $h = f^{-1}$ is the headway between consecutive departures. The final design variable we consider in the analysis is vehicle capacity: $s = \omega Q/f$, in which Q is total hourly demand in both directions of the origin-destination pair, and $\omega > 0.5$ is the share of the busier market in total demand.

Cycle time and service frequency are the two determinants of the minimum fleet size (F) along the line:

$$F = \sigma/h = \sigma f = \left(t_0 + \frac{2d}{v}\right) f. \tag{2}$$

Figure 1 illustrates the feasible combinations of service regularity (measured by the headway between consecutive departures) and travel speed, keeping fleet size F fixed. The visualisation confirms that higher velocity enables more frequent services with a given fleet size, but the incremental headway benefit diminishes in speed as well as frequency itself. Several results in subsequent parts of the analysis will be rooted in these implicit properties of scheduled transport operations.

Speed has numerous impacts on both user and operator costs, and the aim behind the optimisation of velocity is to find the right trade-off between these costs. Let us define a social cost function composed of three additive elements.

$$C(Q, v, f, F, s) = C_k(v, F, s) + C_o(v, f, s) + Q \cdot c_u\left(h(f), t(v)\right) \tag{3}$$

[2] For a more detailed engineering analysis of speed, vehicle dynamics and energy efficiency, see X. Li & Lo (2014).

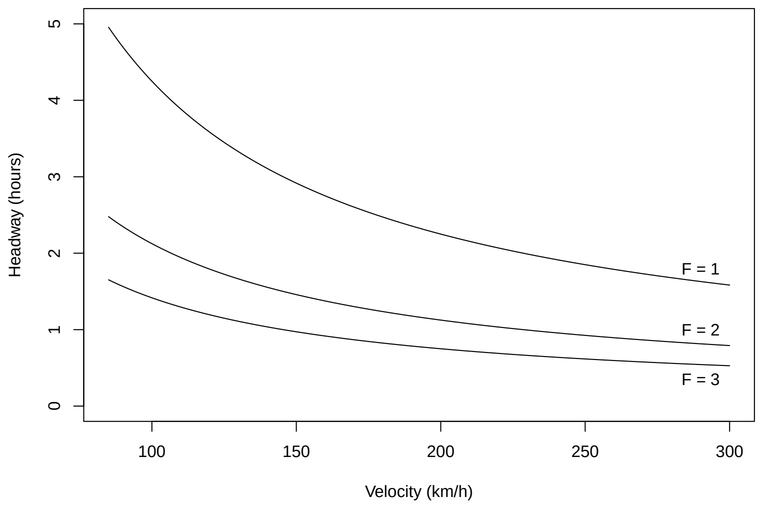

Figure 1: Feasible headway-velocity combinations with a fleet of 1, 2, and 3 trains. Parameters: $d = 200$ km, $\alpha = \$20$ per hour, $\alpha_w = \$10$ per hour, $t_0 = 15$ min.

In equation (3) we indicate the functional dependencies between cost components and the design variables introduced above. $C_k(v, F, s)$ denotes the capital cost of service provision. The full paper includes mathematical formulae; in this short version we describe the specification verbally:

- Capital cost is split between infrastructure provision[3] and fleet management:
  $C_k(v, F, s) = d \cdot \rho(v) + F \cdot c_k(v, s)$
- The operator's expenditure depends on vehicle kms and the speed and size of trains:
  $C_o(v, f, s) = 2df \cdot c_o(v, s)$
- The total user cost is the sum of in-vehicle time and schedule delay cost:
  $c_u(h(f), t(v)) = \alpha\, t(v) + \alpha_s\, h^\beta$,

Note that a trade-off between headway and velocity exists in the user's perception as well. Let us compute isocost curves on the headway-velocity space along which the user cost function remains constant. Figure 2 depicts three such isocost curves for a set of parameters provided in the figure's caption. Relatively slow but frequently scheduled service can sometimes be just as convenient as a very fast but rarely available alternative.

3 Results and discussion

The cost functions defined above are suitable to quantify the cost implications of the choice of speed, frequency and vehicle size, keeping demand (Q) parametric. Even though demand is parametric, the cost minimisation exercise enables us to perform additional analyses of welfare-oriented pricing decisions and the financial performance of service provision, without specifying an explicit demand system. Microeconomic theory suggests that travellers perceive the social surplus maximising monetary incentive when they are required to pay the net marginal social cost of their trip minus the user cost they already bear (Small & Verhoef, 2007; Czerny & Peer, 2023). Let us call this quantity the marginal *non-personal* cost of travelling to distinguish it from the derivative of the *total* social cost, $\partial C/\partial Q$. The marginal non-personal cost is then

$$\frac{\partial C(\cdot)}{\partial Q} - c_u(\cdot) = \frac{\partial C_k(\cdot)}{\partial Q} + \frac{\partial C_o(\cdot)}{\partial Q} + Q\frac{\partial c_u(\cdot)}{\partial Q}, \tag{4}$$

which is the sum of the capital, operational and external user cost induced by the marginal traveller. Another key point of microeconomic theory is that the efficiency-maximising monetary

[3]The cost of infrastructure provision may include new construction as well as the volume of non-expansion capital expenditure on renewals; see an economic model of the latter in Xuto et al. (2021).

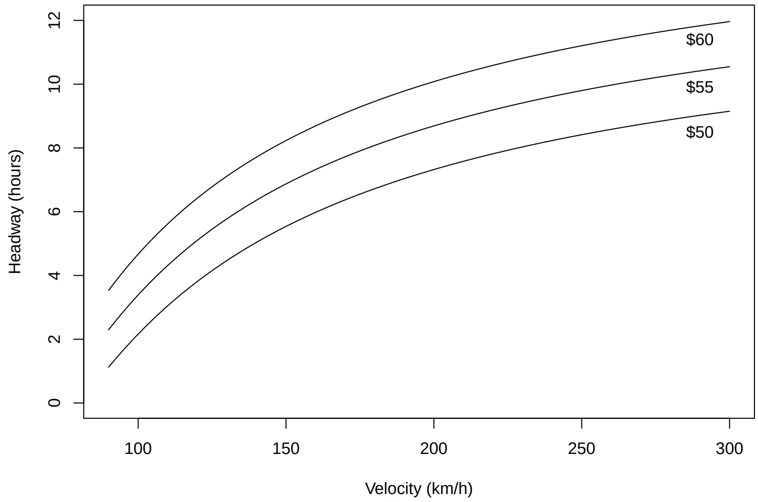

Figure 2: Users' isocost curves in the headway-velocity space. Passengers perceive the same user cost at any combination of service frequency and average travel speed along the isocost curve. Parameters: $d = 200$km, $\alpha = \$20$ per hour, $\alpha_s = \$5$ per hour, $\beta = 0.9$

incentive (fare) is the *short-run* marginal non-person cost, that is, the cost directly attributed to the incremental trip. Let us enlist three alternative assumptions that may affect our analysis fundamentally:

1. Capacity (including speed, frequency and vehicle size) are completely fixed in the short run. In this case capacity shortages emerge in response to marginal changes in demand. If capacity imposes a strict limit in the form of a vertical supply function, the role of pricing is keep demand at the capacity limit. This case might be highly relevant for existing rail services running on existing infrastructure. However, as this case is not relevant from the viewpoint of the optimisation of speed, we do not perform analysis under this assumption.
2. Speed is predetermined along an existing rail line, but the operator is able to adjust service frequencies as part of a short-run timetable reform. Thus, the marginal social cost fare includes the capital and operating cost of marginal frequency adjustment.
3. Assume that we are in the planning phase of a new rail service, such that all capacity variables, including speed, are endogenous in the model. The planner's aim is to set the fare equal to the full expression in (4).
4. As a specific case of assumption 3, assume that service frequencies are predetermined by a network-wide cyclical timetable but a specific railway line is under a short-run infrastructure overhaul. Speed is re-optimised for the estimated level of demand. Thus, the marginal social cost in (4) includes the capital and operating cost of speed adjustment, but frequency adjustment is excluded.

Due to the length limit of this short paper, we provide results for case 3 only which is often referred to as unconstrained *strategic design* (Jara-Díaz et al., 2023).

The first-order condition of the optimal speed yields the following optimality rule:

$$v = \sqrt{\frac{\alpha Q + 2fc_k}{\rho'(v) + \frac{F}{d}\frac{\partial c_k}{\partial v} + 2f\frac{\partial c_o}{\partial v}}}. \tag{5}$$

A key message of this cost-minimising speed rule, formerly not shown in the literature, is that the optimal velocity of a rail service depends on the volume of travellers on the line. This message questions some of the influential policy decisions in the long-distance rail market where public bodies often try to standardise the design speed of new rail infrastructure throughout their jurisdictions. We observe nation-wide regulations in this spirit.[4]

[4]See, for example, the uniformity of speed on the newly built high speed rail networks of China,

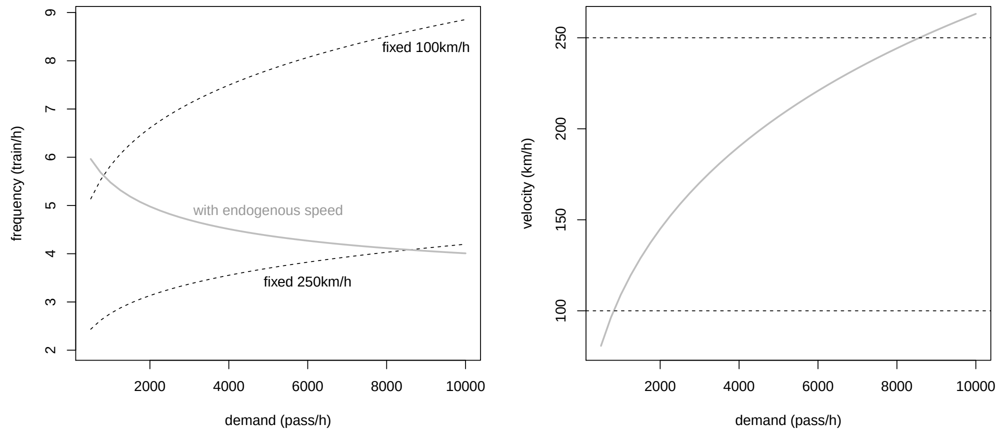

Figure 3: Social cost-minimising capacity policy with endogenous velocity

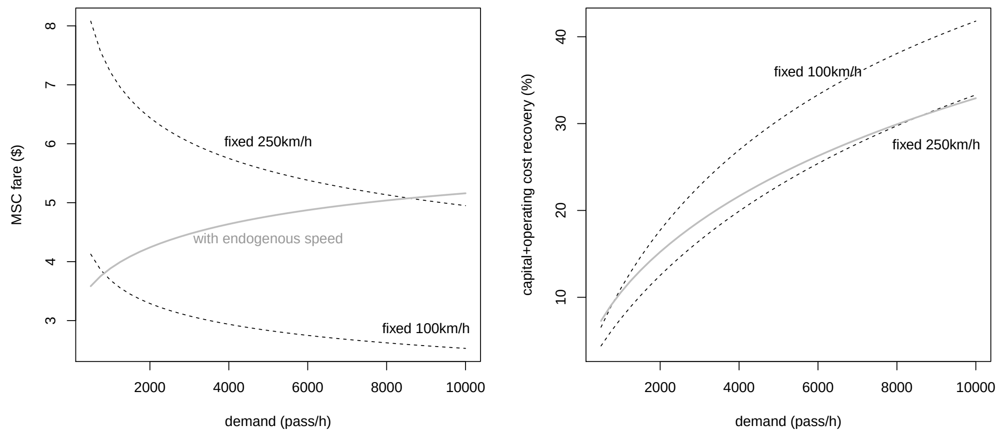

Figure 4: Approximation of the social surplus maximising fare and cost recovery ratios under optimal supply; endogenous velocity

The optimal velocity rule shows close similarity with the previously known versions of the optimal frequency rules in the public transport literature in the sense that the optimum increases with a function of the square root of demand. This is not surprising given that both f and v enter the user cost function in a reciprocal form: waiting time depends on the headway, f^{-1}, while travel time is also inversely related to speed through $t = dv^{-1}$. As a consequence, after taking derivatives, the rearrangement of first-order conditions leads to square-root formulae in both cases. Interestingly, the demand-dependency of frequency has received much more attention in the literature since the seminal contribution of Mohring Mohring (1972), while common-sense intuition suggests that travel speed is not less important in the perception of users, at least in the long-distance rail context. By numerically simulating the speed rule we show in the rest of this section that the optimisation of speed replicated several properties of the well-known Mohring rule, including scale economies in various cost components.

The two panels of Figure 3 visualise the numerical solutions of $min_{v,f} C(v, f, s, F)$. Panel (b) confirms that speed is a positive concave function of demand. With the model parameters and demand interval considered, the optimal speed varies on a wide range. From a policy point of view this implies that in a national railway network in which demand intensity varies due to the unequal spatial distribution of economic activity, it is difficult to justify a fully homogeneous (standardised) speed for rail services.

and Spain in Europe, and the EU-wide regulation that conventional rail reconstruction must enable a commercial speed of 160km/h.

Panel (a) is even more surprising in light of the existing literature of public transport supply. More specifically, Mohring's well-known frequency formula prescribes that frequency should be proportional to the square-root of demand Mohring (1972), and this finding has remained consensual in the past half century Hörcher & Tirachini (2021). Results not included in this short paper indicate that the 'square root principle' does hold in our model as well – as long as we keep velocity fixed. By contrast, when velocity becomes endogenous, the optimal frequency is a decreasing function of demand, meaning that Mohring's famous rule is no longer applicable.

The combination of the optimal speed and frequency variables suggests strong substitution between the two supply-side variables. Higher speed enables the operator to maintain, and even reduce, the average user cost, and realise capital and operating cost savings by operating a smaller fleet of larger vehicles. The precondition of this strategy is to have sufficient demand intensity on the line. In low-demand markets, high speed does not generate sufficient user benefits and large vehicles would imply inefficiently low frequency and long wait times.

In the final part of this section we infer how efficient pricing decisions and profitability are affected by the endogeneity of speed. Panel (a) of Figure 4 shows that the marginal non-personal cost, our proxy for the optimal fare, increases with Q. This suggests that high-demand (and thus high-speed) markets should be more expensive than low-demand counterparts – at least when speed has been set through an endogenous optimisation process. Panel (b) shows that the overall relative financial performance of the service improves with demand, but the improvement in cost recovery is milder with endogenous speed, as compared to fixed-velocity scenarios.

4 Conclusions

In this research we construct a microeconomic model of intercity transport supply in which the speed of service provision appears as an explicit supply-side variable. In a series of supply optimisation exercises, we show analytically and through a numerical example that the socially optimal speed is an increasing (concave) function of the intensity of demand. The paper reflects on the previous literature of mass transit economics by analysing the optimal frequency pattern with responsive velocity: we show that Mohring's well-known 'square-root principle' does not hold in this case and the optimal frequency actually decreases with hourly passenger demand. In practical terms, this means that thick intercity rail markets should be served at higher speed but lower frequency, compared to less densely used lines where the combination of lower speed and higher frequency appears more efficient.

The paper reveals some of the fundamental mechanisms of long-distance rail service provision. The model we develop is simple to keep the results tractable. The analysis neglects various network effects, including fluctuations of demand along a rail line and the dependence of intermediate stops on train speed. We also neglect interactions with heterogeneous rail traffic, including freight flows (S. Li et al., 2023), and the timetabling problem beyond the choice of frequency. Market structure is stylised as we assume a welfare maximising monopolist – the choice of speed might be affected by competitive interactions in a liberalised railway market. Yet another limitation is the static nature of the analysis: railway infrastructure and rolling stock investments last for decades and the pattern of growth in demand over this time span is often uncertain. This implies that supply optimisation based on a short-run demand prediction may underestimate the optimal velocity during the entire life-cycle of the service.

References

Börjesson, M., Fung, C. M., & Proost, S. (2017). Optimal prices and frequencies for buses in Stockholm. *Economics of Transportation*, *9*, 20–36.

Button, K. (2012). Is there any economic justification for high-speed railways in the United States? *Journal of Transport Geography*, *22*(C), 300–302.

Chang, S. K., & Schonfeld, P. M. (1991). Multiple period optimization of bus transit systems. *Transportation Research Part B: Methodological*, *25*(6), 453–478.

Czerny, A. I., & Peer, S. (2023). Transport pricing: Theory and methodologies. In A. Tirachini, D. Hörcher, & E. T. Verhoef (Eds.), *Handbook on Transport Pricing and Financing* (pp. 24–38). Edward Elgar Publishing.

Hörcher, D., & Graham, D. J. (2018). Demand imbalances and multi-period public transport supply. *Transportation Research Part B: Methodological*, *108*, 106–126.

Hörcher, D., & Tirachini, A. (2021). A review of public transport economics. *Economics of Transportation*, *25*, 100196.

Jansson, J. O. (1980). A simple bus line model for optimisation of service frequency and bus size. *Journal of Transport Economics and Policy*, 53–80.

Jara-Díaz, S., & Gschwender, A. (2003). Towards a general microeconomic model for the operation of public transport. *Transport Reviews*, *23*(April 2013), 453–469.

Jara-Díaz, S., Gschwender, A., & Hörcher, D. (2023). Public transport: design, scale, and pricing. In A. Tirachini, D. Hörcher, & E. T. Verhoef (Eds.), *Handbook on Transport Pricing and Financing* (p. 171-189). Edward Elgar Publishing.

Li, S., Zhu, X., Shang, P., Li, T., & Liu, W. (2023). Optimizing a shared freight and passenger high-speed railway system: A multi-commodity flow formulation with benders decomposition solution approach. *Transportation Research Part B: Methodological*, *172*, 1–31.

Li, X., & Lo, H. K. (2014). An energy-efficient scheduling and speed control approach for metro rail operations. *Transportation Research Part B: Methodological*, *64*, 73–89.

Liu, W., Zhang, F., Wang, X., & Tang, Y. (2023). A review of selected transport pricing, funding and financing issues in Asia. In A. Tirachini, D. Hörcher, & E. T. Verhoef (Eds.), *Handbook on Transport Pricing and Financing* (pp. 380–393). Edward Elgar Publishing.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *The American Economic Review*, *62*(4), 591–604.

Small, K. A., & Verhoef, E. T. (2007). *The economics of urban transportation*. Routledge.

Sun, L., Tirachini, A., Axhausen, K. W., Erath, A., & Lee, D.-H. (2014). Models of bus boarding and alighting dynamics. *Transportation Research Part A: Policy and Practice*, *69*, 447–460.

Tirachini, A., & Hensher, D. A. (2011). Bus congestion, optimal infrastructure investment and the choice of a fare collection system in dedicated bus corridors. *Transportation Research Part B: Methodological*, *45*(5), 828–844.

Tirachini, A., Hensher, D. A., & Rose, J. M. (2014). Multimodal pricing and optimal design of urban public transport: The interplay between traffic congestion and bus crowding. *Transportation Research Part B: Methodological*, *61*, 33–54.

Vickerman, R. (2018). Can high-speed rail have a transformative effect on the economy? *Transport Policy*, *62*, 31–37.

Xuto, P., Anderson, R. J., Graham, D. J., & Hörcher, D. (2021). Optimal infrastructure reinvestment in urban rail systems: A dynamic supply optimisation approach. *Transportation Research Part A: Policy and Practice*, *147*, 251–268.