The social cost of rail capacity constraint

María Pérez Herrero

Ph.D. Candidate
Junior scholar
Laboratoire d’Economie des Transports, Université de Lyon 2
Réseau Ferré de France, Paris

Abstract

Capacity shortage is a standard problem in transport sector, and unlike to rail, it has been studied in-depth for roads and air transport. Nonetheless, each industry has strong specificities and it is complex to establish a conventional definition about what capacity shortage is.

The first purpose of this paper is comparing the capacity shortage analysis in the other transport industries, where the notion of congestion is deeply accepted, in order to establish a parallelism with the rail sector. This analysis shows that congestion in rail sector can be analytically compared to road transport. Nevertheless, distinctive features on the rail sector and the whole factors having and influence on capacity shortage will be discussed. The aim of this paper is to provide a few hints for the development of a comprehensive economic theory of railway capacity

Keywords: capacity; congestion; delay; externality; marginal cost; pricing; optimal trade-off; rail transport; resilience; robustness; scarcity; scheduling.

Corresponding author information here. Tel.: +33 153 94 93 19
E-mail address: maria.perez-herrero.cnrs@exterieur.rff.fr
1. Introduction

In 2012, the French rail regulator (Autorité de Régulation des Activités Ferroviaires, ARAF) pointed out congestion as a crucial issue for the French railway system. In order to improve the utilization of the network, the infrastructure manager should develop track access charges and provide efficient economic incentives (statement 2011-002 of the 2nd February 2011).

Rail capacity is a major issue for the French network. Demand for rail transport grows in metropolitan areas and the expansion of rail capacity faces a range of obstacles and financial challenges.

However, rail capacity is a complex issue, with numerous meanings and not a standard definition. The capacity of a railway line depends on how it is used and how different parameters are combined (infrastructure factors, traffic features, operating requirements and management decisions). Line capacity is essentially what the infrastructure managers have to sell as their final service.

Economic literature is quite extensive about road capacity constraints. In railway sector, the capacity shortage has traditionally been considered as the inability for a train operator to obtain the desired train path (scarcity). However, this capacity perception seems restrictive. A lack of capacity can be experienced before scarcity, as unexpected transmitted delays are positively related to the density of traffic (congestion).

The aim of this paper is to provide a few hints for the development of a comprehensive economic theory of railway capacity, which would be a useful tool for the optimisation of timetabling, but also for the implementation of a rail congestion-pricing scheme and capacity investments. This paper describes a work programme which will be developed in further researches.

The structure of the paper is as follows. Section 2 briefly reviews the academic literature dealing with the economics of congestion in transport, and particularly in rail transport. It presents the intuition behind this paper: high traffic density may generate a higher probability of delays. This intuition is confirmed by econometric analyses from Gibson et al. (2002) and Brunel, Marlot and Perez-Herrero, (2013). Section 3 provides an analogy between road and rail congestion. Section 4 describes the distinctive characteristics of rail congestion analysis. At last, section 5 offers concluding remarks.

2. Related literature

Many networks suffer from peak-load demand problems. In general, congestion refers to the existence of limited capacity networks whose demand varies periodically. A wide literature deals with congestion in the road sector, from classic contributions such as Pigou (1920), Walters (1961) or Vickrey (1963) to more recent works from Arnott, De Palma and Lindsey (1993), Chu (1995) or Verhoef (2001). When cars users decide to make an additional trip, they impose additional costs on themselves, on the infrastructure provider and on other users. From an economic perspective, congestion is basically a standard externality problem. Academic literature shows that peak/off-peak pricing is an efficient solution to tackle congestion, and allows users to internalize the external costs generated and reallocate the traffic demand during the day (Vickrey, 1963). Another interesting conclusion is, that given certain circumstances, congestion pricing covers the construction costs of highways (Mohring and Harwitz, 1962, Arnott and al., 1993, Hau, 1998).

As Quinet (1997) pointed out, congestion does not appear only on roads but also in other transportation modes, even where traffic is scheduled in advance. By contrast with road sector, congestion has received less attention in these sectors. A sizeable literature nevertheless has studied airport congestion that happens in the neighbourhood of large airports due to runways or traffic control saturation.

In a seminal paper, Carlin and Park (1970) estimate the marginal cost of delays in New York’s LaGuardia airport. The congestion cost is defined as the additional delay imposed on the following planes in the queue during the busy period. One of the main results is that landing fees are inefficient, because the congestion costs are significantly higher than what is paid by air carriers. This paper explores the possibility of imposing a
congestion toll. A contemporary paper of Levine (1969) also advocates more differentiated fees depending on
the time of the day in order to reflect congestion during peak hours.

A substantial literature follows these seminal articles. Some of them propose to assess empirically the cost of
congestion. This is notably the case of Morrison and Winston (1989). This paper intends to estimate
econometrically the relationship between airport activity and arrival and departure delays using US data. It
clearly exhibits that an increasing level of activity causes an increase in average delays. In other words, when
capacity is used to its fullest, an additional slot increases the probability of delays due to a reduction in the ability
to recover from an incident. Another interesting contribution to this literature is given by De Rus and Nombela
Merchan (2006), which propose a desegregated analysis of airport delays in Madrid Barajas.

Few academic papers consider congestion in rail transport. Some notable exceptions are the High Level Group
on infrastructure charging (Nash, 1999) and papers of Quinet (2003) and Nash and Matthews (2005). These
papers specify the case of pricing railway congestion from a theoretical point of view.

From an empirical point of view, there are even less papers concerning rail congestion in the economic literature.
The researches concerning the congestion study from an empirical approach can be separated in two categories:
simulation angle or econometric angle. In the simulation research literature some papers have studied rail
congestion. These papers have estimated the delays generated by the operational constraints of a railroad
network: delays for meets with opposing rail traffic on single-track lines, and for following and overtaking
slower rail traffic moving in the same direction, for example. This sizeable literature employs both analytical and
simulation-based methods, and study delays and capacity assessment in railroad line networks with specific
configurations following Frank (1966) or Petersen (1974). These kinds of delays are completely internalized by
the infrastructure manager. The developed tools go after the optimal capacity allocation.

A second type of delays is originated by an incident (failure of the rolling stock, failure of the infrastructure,
inadequate behaviour of the crew, etc.). An incident generates delays to the following trains, and, given the
complexity of the network, a lot of trains can be affected, even on different sections of the network. These delays
are obviously unexpected. They increase as capacity utilization rises, because heavy traffic reduces the network
manager’s ability to resolve the incident, and the delay is transmitted to more trains, with a snowballing effect.
This idea is quite familiar in airport economics as stated previously. It is also intuited by the papers of Quinet
(2003) and Nash and Matthews (2005) for rail transport. These kinds of delays are not internalised by the
infrastructure manager and can be estimated considering an econometric approach.

These delays can be measured with an adequate monitoring system. Very few papers have studied this
phenomenon in the economic literature. For instance, it has been empirically studied by the British rail network
(Gibson et al., 2002). In this paper, a regression analysis confirms the existence of a relationship between
capacity utilization and delays. In this study, an exponential form was chosen to estimate for the relationship
between capacity utilization ($C\_\text{util}$) and reactionary delay ($D\_\text{react}$) across the network.

The regression analysis was performed for 24 strategic routes of the British network. The results of the analysis
show that there is a positive relationship between capacity and unexpected transmitted delays. Recently, (Haith et
al. 2013) have proposed an alternative methodology for the British infrastructure manager which concludes that
performance is as much to do with how capacity is used rather than just how much. In other words the
heterogeneity of the timetable is an important factor.

Similarly, an extensive econometric analysis has been conducted for the French railway network, with
comparable results (Brunel, Marlot and Perez-Herrero, 2013). This study focuses on 42 lines of the French
railway network, with 3 measuring points for each line. It shows a positive econometric relationship between
traffic and unreliability rate or the length of delay. According to the line and its features (allowed speed, number
of tracks, signaling…), it shows a positive econometric relationship between traffic and unreliability rate: an
additional train on the line increases the probability of delays, for it and for the other trains. The marginal
congestion cost is made up of a direct effect which is internalized by the supplementary train and of an indirect
effect that generates an external cost on next users.

Another study have recently realised in Sweden in this topic for freight trains. In this study, Krüger (2013)
analyses how train delays are distributed with respect to size, location and time of their occurrence and the link
between capacity usage and expected delay.
This research fits into the continuity of these works.

3. The economic analysis of congestion, from road to railway

Road congestion speed-flow curve is well identified in the economics and transportation engineering literature. As Hau (1992) shows, this engineer curve can be converted to a travel time-flow curve. Using a constant value of time as a shadow price for the representative driver, travel time is then converted to a money basis which yields time cost, called the average variable cost, AVC (Figure 1). The AVC is composed by the road’s variable maintenance cost (proportional to traffic level), by the vehicle operating cost and by the time travel cost.

Following Hau (1992), time cost element is the main responsible for the upward-sloping portion of the AVC curve: when traffic comes close to maximum capacity ($Q^{\text{max}}$), congestion delay has a negative impact on travel time. The AVC curve (before reaching $Q^{\text{max}}$), corresponds to a standard supply curve. When an initial demand function $Q^d$, intersects the AVC curve at point U, a (stable) equilibrium is said to exist at $Q^0$.

Nonetheless, basic price theory says that whenever the AVC rises, it means the marginal cost\(^1\) curve lies above it. The vertical difference between the two cost curves is the marginal external congestion cost: the additional delay that one driver imposes on the rest, and which is not taken into account by the last driver who joins the network. If we consider the previous equilibrium point U, drivers totally ignore in their decision the resulting external cost imposed on next drivers. On the other hand, the equilibrium point Y, associated to an optimal output $Q'$, takes into account the external congestion cost and other variable costs.

---

\(^1\) Marginal cost is obtained as follows: $MC \equiv \Delta C(Q)/\Delta Q = AVC(Q) + Q \cdot \Delta AVC(Q)/\Delta Q$
Optimal charging should consider the additional time that one driver imposes on others. The Pigouvian tax applied to roads is the toll which closes the gap between the marginal cost and average variable cost curves by emitting the correct signal and creating appropriate incentives.

As mentioned in the literature review, some recent articles have studied the econometrical relationship between traffic and delays in the rail sector. Once this relationship has been corroborated, and in order formalise the economic fundamentals of rail congestion, we aim to establish a diagram that could be comparable with the previous road example.

Brunel, Marlot and Perez-Herrero, (2013) proposes a mathematical framework in order to estimate empirically the marginal congestion cost in railways. This mathematical framework enables to calculate the marginal effect of a train on the total delays.

In this section, we notice $R_i^*$ the deviation between the real time and the scheduled time of a train for a given traffic density $Q_i$. The train can be on time ($R_i^* = 0$), early ($R_i^* < 0$), or late ($R_i^* > 0$).

We define the variable $R_i$ representing the delay of train. One can therefore notice:

$$ R_i = \begin{cases} 0 & \text{si } R_i^* \leq 0 \\ R_i & \text{si } R_i^* > 0 \end{cases} $$

The expected delay of train for a given traffic density is:

$$ E(R_i) = p(R_i^* \leq 0) \cdot E(R_i | R_i^* \leq 0) + p(R_i^* > 0) \cdot E(R_i | R_i^* > 0) $$

As the expected delay is null when the train is on time or early ($p(R_i^* \leq 0) \cdot E(R_i | R_i^* \leq 0) = 0$), this equation can be written:

$$ E(R_i) = p(R_i^* > 0) \cdot E(R_i | R_i^* > 0) \quad (1) $$

This equation indicates that the expected delay of a train for a given traffic density is equal to the product of the expected delay of delayed trains and the number of trains delayed.

The total amount delays of trains for a given traffic is, by definition, the expected delay of train multiplied by the number of trains, i.e. $Q_i \cdot E(R_i)$. Therefore, it follows that the marginal delay imposed by an additional train is the derivative of the total amount of delay function with respect to the level of traffic.

It can also be written as:

$$ \frac{\partial [Q_i \cdot E(R_i)]}{\partial Q_i} = Q_i \left[ \frac{\partial E(R_i)}{\partial Q_i} \right] + E(R_i) \quad (2) $$

In this equation, the right hand term is the expected delay of the additional train given the traffic density: this is a direct effect, internalized by the train. The direct effect is equal to the expected delays for a given traffic density. This term, expressed by equation (1), can be directly computed from the data set.

The left hand term of equation (2) represents the marginal delay imposed by the additional train on the following trains. It is an indirect effect which corresponds to the pure externality effect of congestion. The indirect effect, cannot be computed directly and needs and econometrical analysis in order to be estimated.

In fact, the results obtained can be directly compared to the road situation. The MC curve described at the road diagram corresponds to $\frac{\partial [Q_i \cdot E(R_i)]}{\partial Q_i}$ (the addition between the direct and indirect effect) and the AVC curve equals $E(R_i)$. The difference between both curves $Q_i \left[ \frac{\partial E(R_i)}{\partial Q_i} \right]$ (which corresponds to the indirect effect) corresponds to the marginal congestion cost. Using the numerical results for the estimations, we can observe this comparison graphically.
Figure 2 confirms that even in scheduled activities as rail or airports, travel total time is closely connected to traffic flow. Average delay climbs upwards because heavy traffic reduces the network manager’s ability to resolve an unexpected incident, and the delay is transmitted to more trains, with a snowballing effect. As in the road figure, marginal cost curve lies above AVC curve. The vertical difference between both curves is the indirect effect so the marginal external congestion cost: the additional delay that one train imposes on the other trains and that is not internalized.

4. Rail congestion distinctive features

Despite the parallels between road and rail congestion described above, rail sector presents special features that must be taken into account. This analysis cannot be isolated from others parameters in order to have an entire outlook on rail congestion concept.

4.1. Rail transport is a guided transport, which means that train paths are programmed in advance. Compared to road, where vehicles can enter randomly, one can imagine that congestion is internalized in the scheduling process. Nonetheless, the econometric analyses described before shows that it is not completely the case and that we can observe a residual congestion.

Nowadays, schedule is built under technical criteria and an implicit choice is done between the robustness of the schedule and capacity. A detailed analysis of the scheduling process should allow us to understand the total travel time considered for each train path.

First of all, a standard travel time (net travel time) is designed depending on various operational aspects. The first aspect is the track configuration. The network can consist of single, double, triple or even more track, there can be more or less junctions, and the signalling system can allow for more or less trains. The second aspect is the characteristic of each train, such as train length, speed, acceleration rate and deceleration rate (they need to be considered in order to increase or reduce speed without violating the speed limit), and priority (to cross a junction or seize a track, the train with the lower
priority should wait and stop until the train with higher priority passes). The third operational aspect is the speed limits on the different track segments and junctions. Sometimes trains cannot be dispatched at their maximum speed; different trains can have different speeds limits even though their paths may use the same tracks, etc.

Secondly, a recovery time allowing recuperating from an incident or some rail works is taken into account. In France, the recovery time is different between the intercity and regional lines and the high speed lines. For the first ones, an extra time of 4.5 minutes is added for each 100 km. For the high speed lines (HSL), the recovery time equals 5% the standard travel time. Finally, some extra time can also be added for specific rail works or exceptional situations during a bounded time.

Furthermore, with the purpose of minimizing irregularity rates and improving robustness, some extra times, higher than the standard recovery time are added to the schedule. The first possibility is adding an extra interval to the block time (the security interval between two trains). For example, in a HSL, is it possible to run a train every four minutes, but in some lines, this interval equals five minutes in order to add an extra time for recovering from risk. The second possibility for increasing robustness is adding an empty train path between some trains. For example, after three slots with HS trains, the fourth one is empty.

These kinds of practices have as purpose to assure the robustness of the schedule (they minimize the probability to be delayed for the passengers) but they also reduce capacity that could be potentially expended for running extra trains.

In fact, the infrastructure manager internalizes the expected delays when he designs the path of trains. This does not signify the absence of delays in the network. Residual delays appear since we observe a second type of delays: unexpected delays.

Our analysis of the scheduling process shows that scarcity and congestion are strongly linked and implicitly there is a trade-off between capacity (i.e. the number of trains allowed to access the network), and delay. The smaller is the slack time, the greater is the number of paths offered, but also, the lower is the resilience of the timetable, and the greater is the probability of delays: on the one hand, large headways reduce the transmission of delays, but on the other hand, they also reduce capacity. The slack time can be very significant: it represents 20% of the capacity on the French high-speed lines, for instance.

The fact that rail is a scheduled transport has two consequences in the congestion analysis:

i. We cannot observe the “real” capacity shortage, as the scheduling process is a trade-off between capacity and delay. The observed delay at our analyses and scarcity are the result of the implicit choices made at the scheduling process. The impact on delay due to an extra train could be different with different scheduling process rules. It raises the followings questions: How can it be modelled? How is it possible to measure the “non-observed” congestion? Fig. 4 shows the
snowball effect of a delay on following trains. In fact, the difficulty on modelling congestion on rail transport is that this effect is not absolute in itself; it is relative to the scheduling process rules (based on technical and commercial criteria) and its measured can be considered biased.

![Fig.4: Resilience mechanisms in rail roads. (Source SNCF)](image)

**ii.** If congestion is the result of a trade-off between capacity and delays in the scheduling process, then one can expect that, from a long-term point of view, the optimal level of congestion is the result of an adequate investments policy. The infrastructure manager can reduce the congestion when he invests in additional capacities. In particular, one may believe that a benevolent infrastructure manager should expand capacity until the marginal benefit equals the marginal cost of building it. On the contrary, the infrastructure manager could limit the capacity or the number of paths offered, in order to raise unjustified profits by limiting the capacity.

Furthermore, the literature also demonstrates that, given certain circumstances, congestion pricing covers the construction costs of infrastructure (Mohring and Harwitz, 1962, Arnott et al., 1993). Hau (1998) proposes a rigorous non-mathematical interpretation of this literature and relaxes some of the assumptions of the cost recovery problem. It includes the constant return of scale and the perfect divisibility of the investment hypotheses.

This is why the natural monopoly must be regulated. The regulator must compare the congestion cost (and, if needed, the scarcity cost) to the cost of an increase in capacity. It must also assess the optimality of scheduling through a cost benefit analysis.

From the infrastructure point of view, scheduling and capacity investments are only different ways to tackle a capacity shortage, short term and long term. One must be aware that, if the scheduling process internalizes some of the congestion, which should not be taken into account in the optimal track access charges, the value of this internalized congestion should be taken into account in the assessment of capacity investments.

4.2. In railways, there exists a separation of infrastructure from operational services. Upstream market (infrastructure) is a natural monopoly. On the other hand, downstream market (operational services) can be in monopoly (under imperfect regulation) or in competition, depending on the activities. Under these conditions, it is difficult to model and simulate the effect of an optimal congestion toll. How the price signal on infrastructure train paths would be transmitted to passengers? The analysis of an optimal congestion toll cannot be dissociated to a global market analysis and regulation context. How the downstream market is organized? Should be a difference in the analysis approach when activities are purely commercial or they are public service obligation? Nowadays, there exists a competition from the network utilization between different operators and different activities (freight, regional transport authority, SNCF, Eurostar, Thello, etc.), how that should be considered in the congestion analysis? The same train company operates different activities with different organizing authorities; it means that the choice of a determinate train path is made by different planners. In that case, can we assume that the TOC internalize the congestion cost? Or do we expect that as they are in competition for some paths, the delay cost imposed on others train is not completely taken into consideration.
5. Conclusion

The present paper proposes an economic analysis of congestion in rail transport. It shows that the economics of congestion in rail transport can be compared to road transport analysis. As traffic is scheduled by advance, one can expect scarcity rather than congestion in rail transport. Nevertheless, congestion can appears even when traffic is scheduled in advance: when rail network is highly used, an additional train path increases the probability of delays due to a reduction in the ability to recover from an incident.

This paper formalise the congestion analysis in rail transport establishing a parallelism with road long-established studies. It confirms the existence of an external congestion cost on rail transport.

Nevertheless, rail sector presents singular characteristics (particular market structure and guided transport) that must not be dismissed if we would aim to have a global and realistic analysis of rail congestion phenomenon.

The analytical solution to these special features reflexions remains to be defined and will be developed in further researches.

References


