AN EXPLORATION OF RAILWAY CONGESTION: A PROPOSED METHOD FOR ESTIMATING THE INFRASTRUCTURE UTILIZATION RATE

LAROCHE, Florent, Transport Economics Laboratory (LET), University of Lyon

Florent.Laroche@let.ish-lyon.cnrs.fr

ABSTRACT

This paper proposes a method for estimating the utilization rate of infrastructure. It follows on from work that was carried out to forecast rail transport demand in France in 2050. The utilization rate is defined as the ratio between demand for railway services and the available capacity of the infrastructure. Our findings show that the infrastructure utilization rate varies according to how capacity is defined and that the quality of service varies according to the utilization rate.

Keywords: infrastructure, capacity, railway, utilization rate

INTRODUCTION

This paper considers the relationship between rail transport demand and infrastructure capacity. It originates in a study that was carried out to forecast CO2 emissions in France in 2050 (ENERDATA-LET, 2012). Three scenarios that express changes in transport demand were specified for this study. They are based on macroeconomic and microeconomic hypotheses (respectively, changes in GDP and travel time budgets). These scenarios have two points in common. First, overall transport demand is expected to continue to rise until 2050, unless we consider a long period of negative growth, as transport demand is positively correlated with GDP (Schafer, 2001). Second, modal transfer is expected to be very much to the advantage of rail as a result of two parameters: the elasticity of speed to GDP is positive (Schafer and Victor, 2000; Crozet and Joly, 2004) and the existence of increasingly strong public policy measures to reduce speeds on the roads and encourage high speed rail.

Thus, in France, interurban rail passenger transport is expected to increase by a factor of 3 and more between now and 2050 and rail freight transport by a factor of 1.5 and more, depending on the scenario. This raises the following question: how well could the existing network stand up to such an increase in demand?

This paper proposes a method for evaluating the utilization rate for a piece of railway infrastructure. This is determined on the basis of the interaction between demand and infrastructure capacity. Demand is considered to be fully satisfied. The rail transport production system must therefore adapt, possibly revealing a need for additional capacity. The method we have applied uses revisable parameters to estimate how the utilization rate of the infrastructure affects its capacity and the resulting quality of service.

Section 1 presents the concept of capacity by means of a literature review. We shall then describe the terms of our method before outlining its structure in section 3. Section 4 discusses the theoretical results before the presentation of an empirical application of the method.

SECTION 1: LITERATURE REVIEW

The literature that examines the concept of railway infrastructure capacity (single or double track) dates back some way (Lai, 2008). However, no standard model exists as it does for road transport (Abril et. al, 2008). Each analysis generally uses a specific model, with its own strengths and weaknesses (Martland and Hutt, 2005). We can identify two types of approach. One is a conventional engineering approach that sets out to make a technical evaluation of capacity and the other is a more recent economic approach that deals with the distribution and pricing of the technical capacity.

Until recently the railway system, both in Europe and elsewhere, was characterized by the absence of economic free play. This partly explains why the engineering approach has provided the oldest and fullest body of literature. The studies involved were mainly conducted for railway companies since the 1960s and essentially dealt with identifying the determinants of the capacity of a single or double track. This research is based on theoretical, parametric or simulation models (Abril et al., 2008; Lai and Barkan, 2009).

The first work was conducted in the 1960s (Lach and Skelton, 1968). It was motivated by the need to improve the supply of rail transport in order to respond to demand for higher standards and improved quality (Petersen, 1974). Building on research into road traffic flow (Carleson, 1957; Frank, 1963), Petersen (1974) proposed a probabilistic technique for simulating train journey times on a single track line. This author identified two decisive factors, namely the heterogeneous nature of the speeds of the different railway services and the system of priority, which depends on this. In order to make a more comprehensive estimation of capacity, Prokopy and Rubin (1975) developed the first parametric model. Kraft (1987, 1988) added an algorithm to estimate the impact on train movements and delays of the decisions made by railway managers with regard to the hierarchization of flows. He also quantified the distinction between the “practical” and the “theoretical” capacity of a railway line (1982), estimating the former as 60 - 75% of the latter. This distinction allows brings in the key concept of system flexibility (or network resilience) for a scheduled activity. According to the definition given by the SÉTRA (2009), the “margin of flexibility” makes it possible “to avoid saturation and knock-on delays in the event of incidents”.

The resilience of a network relates to its ability to return to normal operation after a disruption. Higgins et al. (1995) formalized the above work in mathematical language and included the principle of optimal scheduling according to the level of train dispatching priority. Krueger (1999) developed a novel parametric model in which the determinants of capacity (infrastructure, traffic management and operating rules) were described with a view to proposing an optimization. This research has been summarized in a doctoral thesis by Lai (2008) who studied in greater depth the impact of infrastructure management (operation and planning) on capacity. In a related paper, (Lai and Barkan, 2009) improved Krueger's parametric model (1999) by adopting a dynamic approach towards capacity. This work enriched the model by adding a module that considered investment choices for improving infrastructure capacity and, in particular, modelled how investment can alter the available capacity of infrastructure. In response to this growth in the literature and modelling, the UIC (2004) published the UIC 406 method to harmonize the measurement of railway infrastructure capacity. However, Landex et al. (2006) have shown that it is difficult to define a universally applicable system as each railway system has its specific characteristics. Last, as part of a doctoral thesis, Harrod (2007) proposed a full literature survey and an optimization model intended for infrastructure managers.

The years 2000 saw the emergence of a literature that was more concerned with identifying and describing the determinants of capacity. Based on the work of Kraft (1982), Burdett and Kozan (2006) proposed a straightforward method for defining the "absolute" (theoretical) capacity of infrastructure while Liotta et al. (2009) developed a method for evaluating the "commercial" (practical) capacity of a railway line. In addition, Landex and Kaas (2005) examined the buffer time on a highly trafficked line and the relationship between speed and capacity. They showed that the buffer time between two trains is a determinant of capacity and quality of service. The longer the buffer time between two trains, the greater the margin trains have to make up for delays and the more resilient the network becomes. Abril et al. (2008) tested the effect of other parameters on infrastructure capacity. In particular, they estimated the impact of train heterogeneity, the number of commercial stops made by the train and speed disparities on infrastructure capacity and the robustness of timetables. With a view to improving train reliability, D'Ariano et al. (2008) have proposed a real time conflict resolution method and claim that their flexible timetables and the abolition of priority would improve system efficiency. Last, more recently, research has considered the determinants of delay and examined capacity on the basis of the quality of service concept (Liotta et al., 2009; Dingler et al., 2010; Schlake et al. 2011). This approach is based on earlier work by Higgins et al. (1995) which evaluated the determinants of delays and their impacts on timetable reliability.

Economists became interested in the issue of capacity somewhat later. For a long time railway systems consisted of vertically integrated monopolies that regulated themselves, giving priority to supply over demand. Sweden's deregulation of its railway system which got under way in 1989 and was taken up by the European Commission in 1991 (Directive 91/440/EC) opened the way for economic analysis. A shift therefore occurred from a supply-based approach to a demand-based approach under network capacity constraint. The study of the determinants of capacity thus became the study of the value of a train slot with reference to the load on the infrastructure (Harker and Hong, 1994). The vertical separation between the infrastructure manager and the network operator meant the system ceased to regulate itself on its own and information asymmetries (Ferreira, 1997). As a result of this state of affairs, the literature focuses on two research issues: what mode of capacity allocation mode should be applied in order to ensure that network access is not affected by discrimination (Nilsson, 1999, 2002; Gibson, 2003; Nash and Matthews, 2005) and what train slot scarcity charge should be implemented in order to optimise the manager's revenue (Quinet, 2003; Johnson and Nash, 2008; Marlot, 2012)?

This literature review reveals that the supply-based approach to capacity is dominant. The engineering approach sets out to identify the technical factors that determine infrastructure capacity and the economic approach evaluates the cost of access to the infrastructure under capacity constraint. With regard to capacity, a distinction is made between theoretical capacity, which is defined by engineers, and practical capacity, which is the capacity that is made possible as a result of choice of network manager between service quality and traffic demand. This study proposes to link the engineering to economic analysis of railway system. We develop an inductive method, using as our point of reference current and future demand in order to identify infrastructure capacity and any needs for additional network capacity. It is to this end that we shall propose a straightforward theoretical methodology for modelling the capacity of a network and defining its utilization rate.

SECTION 2: THE TERMS OF THE MODEL

Our modelling of infrastructure capacity uses five principal factors in two categories: technical factors (engineering side) and economic factors (results from stakeholders choice and strategy). These factors constrain supply and also constitute adjustment variables according to the level of demand.

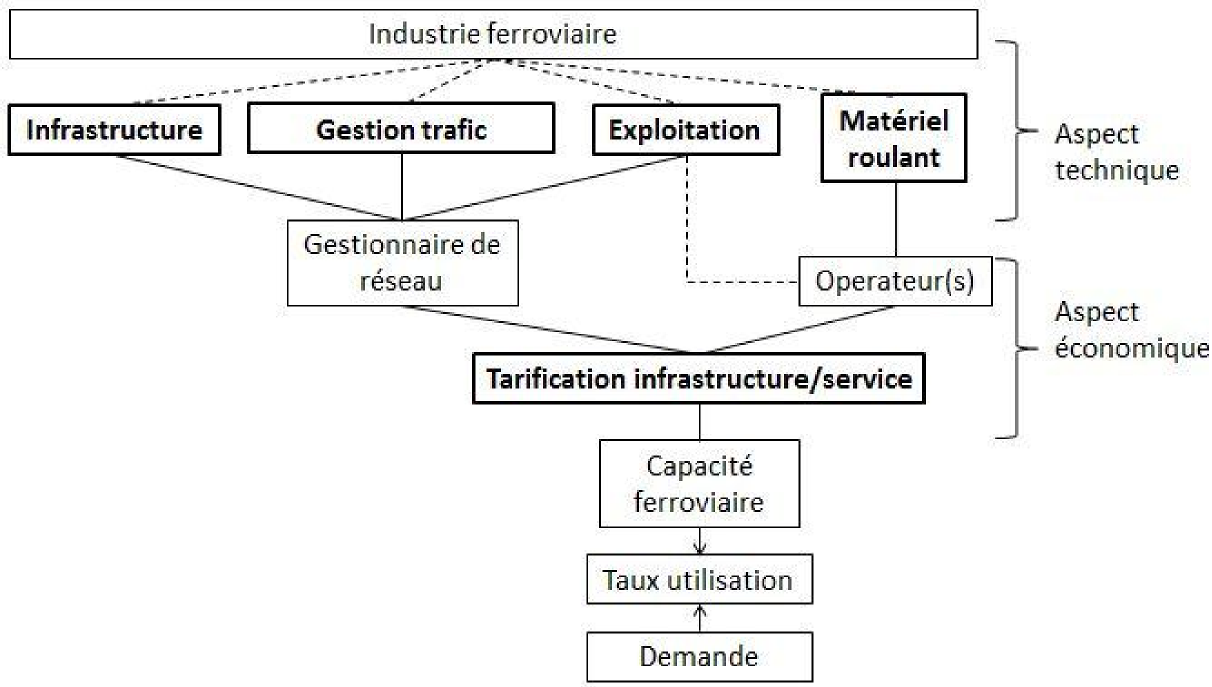


Figure 1: simplified diagram of railway capacity (source: author)

The infrastructure is defined on the basis of the types of rail services it is intended to carry. This characteristic is not unimportant, particularly in France where the development of High Speed Lines (HSLs) makes some traditional services impossible. Three types of infrastructure are identified. HSLs that only carry long distance high speed passenger trains (TGVs). Conventional lines that carry both long distance passenger or freight trains and short distance passenger trains. Last, freight lines that mostly consist of old conventional lines that have been reclassified for freight and short distance passenger trains after an HSL has been built alongside them. The existence of these diverse types of lines, which is specific to France, means that capacity can vary enormously from one line to another. The architecture of an HSL permits higher speeds than a conventional line, but does not necessarily increase theoretical capacity because of the constraints associated with speed (Landex and Kaas, 2005; Harrod, 2009). In this situation, capacity depends on the composition of traffic, HSLs having perfect train type homogeneity (a single type of train service) and the conventional line having the greatest heterogeneity and hence the lowest capacity.

The second principal factor, operation, is an essential component of the capacity of a piece of infrastructure. It is this which determines its “theoretical capacity” (Burdett and Kozan, 2006; SÉTRA, 2009). The “theoretical” capacity “is the theoretical maximum upper boundary of capacity. It assumes all trains are the same, with the same train consist, equal priority, and are evenly spaced throughout the day with no disruptions” (Krueger, 1999). The theoretical capacity can be determined from the performance of the signalling system on the track. A distinction is made between the theoretical capacity and the “practical” capacity which depends on network management decisions. Our model considers the theoretical capacity, to which we apply a penalty coefficient in order to determine the infrastructure’s practical capacity (SÉTRA, 2009). This coefficient expresses the network management constraints that are described later in this paper.

The third principal factor, management, mainly relates to specifying the “flexibility” coefficient which determines the “practical” capacity of infrastructure. Liotta et al. (2009) prefer to use the term “commercial capacity” and highlight the importance of the decisions made by infrastructure managers. The classical trade-off is between traffic density (the number of trains per hour) and the required level of service, as the risk of delays increases with the infrastructure’s rate of utilization (Gibson, 2002; Marlot, 2012). This coefficient brings together several dimensions that are characteristic of the decisions made by infrastructure managers. It expresses the need to carry out maintenance works which may remove some capacity. In addition, it represents the available capacity that the manager keeps in reserve for use in the event of disruptions. Last, it expresses the resilience of the network, i.e. its ability to return to normal operation after a disruption. This concept involves both technical aspects and more qualitative parameters such as the speed with which information travels within and between stakeholders, their level of dialogue, experience and trust, etc. Consequently, characterizing a network’s coefficient of flexibility involves a partial evaluation of the railway system as a whole.

The fourth principal factor relates to the rolling stock and its capacity. We make the hypothesis that its technical characteristics are covered under operation and infrastructure (braking time, speed, weight, etc.). Using the capacity of the rolling stock as a variable therefore provides a way of representing some of the decisions made by the operators. It also provides a way of measuring how operators’ decisions and the rolling stock that is industrially available affect railway services in general, because not only the operators, but also industry (which produces the rolling stock) play a role in determining the capacity of a train set. This approach allows us to avoid the issue of the total passenger capacity of the line and consider the capacity of an individual train (Marlot, 2012). The operator thus makes similar trade-offs to the infrastructure manager, namely choosing between the passenger load factor of a train set and the quality of service.

The five and last principal factor relates to the stakeholders choice and strategy. According to the technical constraints and opportunities, each stakeholder defines its strategy. The network manager give a price to the infrastructure use (marginal cost, full cost) and the operator give a price for its service. In general, it depends on demand according to constrain capacity. The consideration of pricing strategy in the model allows to considering the choice made by stakeholders between a maximisation of user’s surplus and a pricing to the marginal cost for a maximisation of demand and social utility of infrastructure and service.

Our method is based on these five principal factors which we consider determine capacity and utilization rate.

SECTION 3: THE STRUCTURE OF THE MODEL

We have defined the infrastructure utilization rate taking as our starting point our data, which are an expression of demand. When following this inductive approach, we decided to develop a specific methodology rather than to use an existing one. However, our work has been guided by the previous work and findings that we have presented in the literature survey.

The data provided by the ENERDATA-LET study (2012) is expressed in passenger-kilometres and tonne-kilometres of travel on the French network. It is distributed between 12 lines, each consisting of one or two routes at most. As this data is in x-kilometres, it is important to have a precise idea of the dimensions of the infrastructure in question in order to obtain a reasonable representation of demand in terms of passengers or tonnes of freight. By converting the data into passengers or tonnes per train we can simulate the commercial strategies of the operators with regard to infrastructure capacity. The variable we selected is rolling stock and we have defined its capacity by the number of tonnes or passengers carried. In order to identify the commercial policy implemented by the operator, we can make a distinction between the theoretical capacity of a train set (its maximum number of tonnes or passengers) and the practical capacity which depends on the percentage of the theoretical capacity which is used. We can consider that the choice of rolling stock and carrying capacity will depend on the strategies of the operators, as the capacity of the network will depend in part on the equilibrium between the capacity of the rolling stock and the frequency of services (Schlake et al., 2011).

This data relates solely to inter-regional and international flows and ignores regional traffic. To make up for this shortcoming we have applied a capacity utilization coefficient to the lines that carry regional traffic. This additional variable allows us to measure how much capacity slower short distance traffic takes away from long distance traffic with lower stopping frequencies and higher speeds. Dingler et al. (2009) have shown that there is a negative correlation between the variety of services and available capacity.

Our proposed method is in three stages. First, demand is modelled by converting tonne- and passenger-kilometres of travel into trains per hour. Second, supply is defined in terms of available slots per hour on the basis of an analysis of the theoretical and practical capacity. Last, demand is compared to supply in order to obtain an infrastructure utilization rate.

Stage 1 conversion of passenger-kilometres of travel (PKT) into trains per hour on each line (TLph):

The first stage sets out to convert passenger and tonne kilometres of travel (PKT) into trains per hour (TL_{ph}).

Three essential parameters are considered: the length of the network, the theoretical capacity per train in terms of passengers and tonnes of freight weighted by the average load factor and the number of hours the network operates. Freight is converted separately from the passengers. We obtain the following equation for passenger and freight traffic:

$TL_{ph} = [(PKT/(Q_t*\theta))/ R]/ H_p$ (1)
Where:
R is the length in kilometres of the passenger routes on each line;
Q_t is the theoretical capacity of a train in terms of passengers;
$0 < \Theta < 1$, θ is the average load factor of passenger trains;
H_p is the number of hours in a year the transport service operates for.

From the average number of passengers or tonnes of freight per train we can deduce the number of trains running in a year which can be converted into the hourly number of trains on the basis of the number of hours the infrastructure operates every day.

This gives us a uniform number of trains per hour per line. In stage 3, we introduce a traffic concentration coefficient which varies according to the type of traffic (freight or passenger), in order to generate heterogeneity in the distribution of trains over 24 hours and simulate peak periods which determine the maximum capacity that is required.

Stage 2: Estimation of the capacity of a line (C_P)

The aim of this stage is to determine the practical capacity of a line (C_P). As we have seen above, this is determined, as stated in SÉTRA (2009), by a coefficient of “flexibility” which reduces the theoretical capacity of the infrastructure.

$C_P = C_T * k$ (2)
Where:
C_P denotes the “practical capacity” of the infrastructure as opposed to its “theoretical capacity” (C_T);
$0 < k < 1$, where k is the coefficient of flexibility.

Unlike most of the models described in the literature that set out to define infrastructure capacity, we do not make any attempt to calculate the theoretical capacity mathematically. We consider that the theoretical capacity is given by the operating system that is applied on the line. It is therefore equivalent to the maximum number of available slots the adopted operating system permits on a line. The operating constraints are represented by the coefficient of “flexibility” k whose value lies within the interval [0 – 1]. To obtain the theoretical capacity of a line, the value of k must be equal to 1.

This equation applies to all types of line, but requires good knowledge of the operating system. In our method, a line can be made up of one or two routes. The capacity of each of these should therefore be calculated on the basis of their characteristics.

Stage 3: Definition of the rate of infrastructure utilization (C_o)

The rate of utilization is the simple ratio between demand and supply. Demand is formulated in terms of trains per hour and supply on the basis of the number of available slots.

$C_o = [((Tl_{fh} * \lambda_f * \varepsilon) + (Ta_{ph} * \theta_p)) / (C_{p1} + C_{p2})] * 100$ (3) in the case of two routes

$C_o = [((Tl_{fh} * \lambda_f * \varepsilon) + (Ta_{ph} * \theta_p)) / C_p] * 100$ in the case of one route

Where:

λ_f et θ_p are the coefficients of concentration, for freight traffic and passenger traffic respectively.

ε is penalty coefficient for freight traffic.

With regard to demand, a distinction is made between passenger and freight traffic. This makes it possible to assign what we have termed a “concentration” coefficient to each. This coefficient, which we have taken from a methodology developed to evaluate the occupancy rate of a road (SÉTRA 2009), allows us to simulate the peak period. The ε is a coefficient to simulate the heterogeneity of traffic mostly between freight and passenger services.

Here, we are considering a line. Consequently, we decided to sum the capacities of the various possible routes. This gives the average utilization rate for the line. It is fairly easy to disaggregate the data according to the route, particularly in the case of France where the presence of two routes often means they are specialized, one carrying in freight traffic and the other mainline passenger traffic.

This method provides a straightforward way of estimating the utilization rate of railway infrastructure. Our demand-based approach to capacity means this model takes account of all the variables that are critical for the railway system, from those that relate to the operators’ strategy under demand constraint to that of the infrastructure manager under capacity constraint. In both cases, our approach permits a trade-off between quality of service and capacity according to the utilization rate, which we shall now evaluate.

SECTION 4: AN ANALYSIS OF THE RATE OF UTILIZATION

The rate of utilization may be defined with respect to either the theoretical capacity (Ct) or the practical capacity (Cp). Our method allows both estimates to be made, depending on the values used. The graph below shows how important it is to clearly define the reference capacity (C) in order to have an accurate interpretation of the rate of utilization

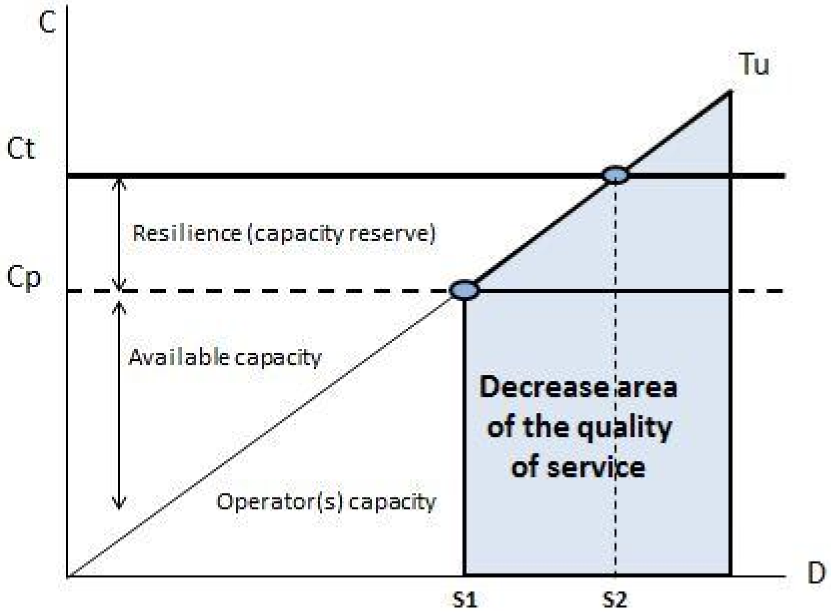


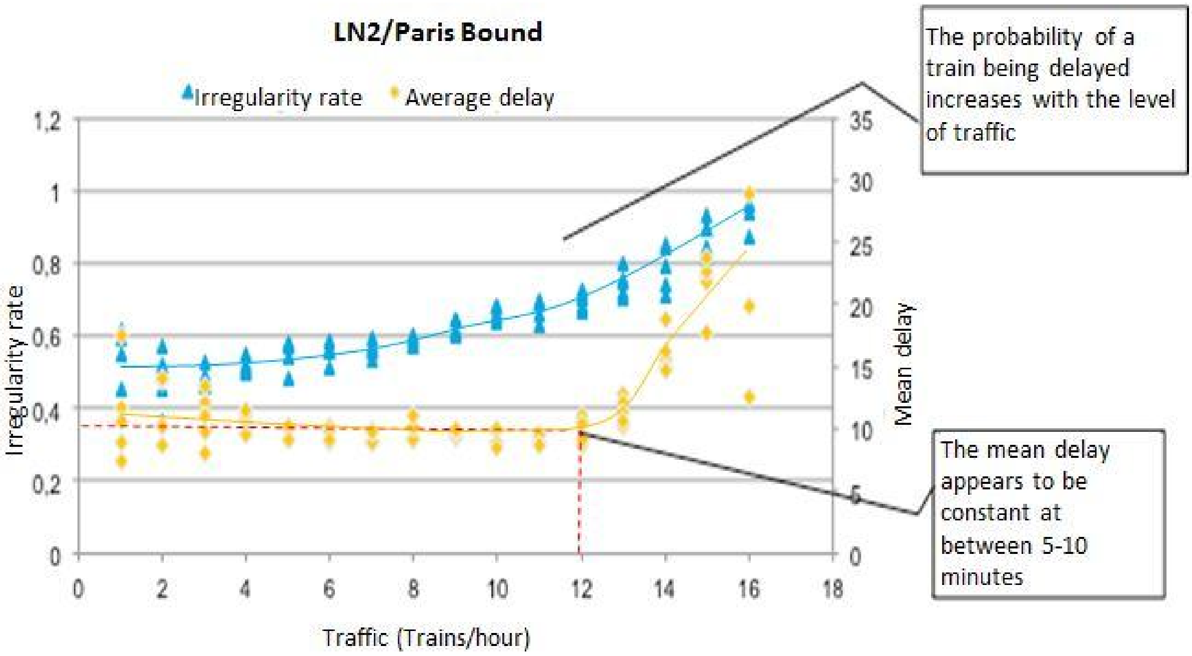
Figure 3: Different interpretations of the rate of utilization (Tu) (source: author)

We can make one initial general observation. The higher the traffic density (D) on infrastructure, the higher the utilization rate and the lower the capacity. A utilization rate that is higher than 100% of theoretical capacity is therefore synonymous with complete saturation. The practical or commercial capacity (Liotta et al., 2009) takes account of the quality of service the manager seeks to provide its customers (the operators). The difference between practical and theoretical capacity corresponds to the coefficient of flexibility that is introduced in our method. This coefficient includes the time required for maintaining the network and the level of resilience expected by the manager in the event of disruptions on the network. To simplify, the diagram posits a direct opposition between resilience to risk and the risk of delay. The resilience is determined on the basis of the “buffer time” specified by the infrastructure manager between each service (Landex and Kass, 2005). It can be seen that the greater this is, the lower the risk of delay but the lower the available commercial capacity. On the other hand, the lower it is, the higher the practical capacity and the higher the risk of delay in the event of a disruption.

We can identify two intersections of the line for the rate of utilization, which give two thresholds. The first (S1) depends directly on the decisions made by the infrastructure manager, which are dependent on the level of service that is provided. However, if social pressure continues to rise, the first threshold can draw closer to the second (S2) at the risk of saturating the infrastructure. It is in the manager’s interest to augment the theoretical capacity of its infrastructure if and only if demand continues to increase. Thus, the first threshold depends on the traffic management policy implemented by the infrastructure manager while the second calls for a modification in the factors that determine the capacity of the infrastructure, i.e. the operating rules, the rolling stock or the infrastructure (new track, etc.).

According to the literature, the practical capacity of the infrastructure is generally approximately 70% of its theoretical capacity (Kraft, 1982; Association of American Railroads, 2007; SÉTRA, 2009). This percentage may vary depending on the railway traffic conditions that apply during peak or off-peak periods. Commercial capacity is often higher during peak periods than during off-peak periods. As demand is greater during peak periods, the social value of the network is increased. This puts pressure on the manager to provide additional train slots, which automatically reduces supply. We can thus observe a reduction in capacity – which economists refer to as congestion – which the infrastructure manager can use to justify higher prices during peak periods. The manager's aim is to capture the operators' surplus in order to maximize its revenue while maintaining a given level of service.

The graph below, which was produced for the manager of the French railway network (Marlot, 2012), provides an empirical illustration of the correlation between traffic density and deterioration in the quality of service. The measurements were made on infrastructure which is known to be congested. In general, the graph confirms the hypothesis that the risk of delay increases with traffic density. However, a more detailed examination reveals a contrasting situation. The "rate of irregularity" curve increases proportionally to the rate of utilization while the "mean delay" curve exhibits a high degree of elasticity above a certain level of traffic. This level is 12 trains per hour, i.e. about 70% of the rate of utilization of the infrastructure with the running of an additional train increasing mean delay quasi-exponentially.



The probability of a train being delayed increases with the level of traffic

The mean delay appears to be constant at between 5-10 minutes

Figure 4: An empirical approach to the ratio between the number of trains per hour, the rate of irregularity and the mean delay (source: Marlot, 2012)

On the basis of this graph, the average delay seems to be the variable that best expresses the difference between practical and theoretical capacity. Above a certain threshold, the introduction of an additional service has a significant impact on the quality of service. We can therefore assume that the resilience of the system is reduced, transforming the slightest delay into a major disruption as the result of a chain reaction. It is there necessary to make a distinction between the effect of scarcity on a network that is expressed by the rate of irregularity and the quality of service which is expressed by the mean delay, which is intrinsically linked to the resilience of the system. Increasing scarcity on a network is due to physical laws while the quality of service seems to depend on the ability of the infrastructure manager to manage traffic and operate services.

To finish on this point, we shall propose an interpretation table for the rate of utilization and its impacts on quality of service. For this we shall refer to two studies. The first provides a threshold-based interpretation of the quality of service based on the utilization rate (Association of American Railroads, 2007). The second is concerned with highway congestion and proposes an indicator that we shall apply to railways. This indicator (the "constrained time") is that time during which "a vehicle is prevented from moving freely, more precisely when it is constrained by the vehicle in front […]. The constrained time is the time light vehicles spend travelling slowly" (SÉTRA, 2009). Translated into the language of railways, this indicator can represent the mean delay. In contrast to what is the case for road transport, in rail transport, the mean delay is not real. As rail activities are scheduled, it is a risk that is associated with the rate of utilization and the resilience of the system. We could therefore define the constrained time in the case of rail transport as follows: "a train is considered to be constrained if its risk of delay is increased as the result of a disruption. This risk is responsible for the mean delay whose level depends on the rate of utilization and the resilience of the system".

Table I – Interpretation table for the rate of utilization (source: author according RFF, 2012 and Association of American Railroads (2007)

Rate of utilization (%)	Impact on the quality of service	Risk of delay	Mean delay
0 - 70	**Guaranteed quality of service** with a sufficient capacity reserve for maintenance and disruptions	Low	< 10 min
70 - 80	**Impaired quality of service** with a moderate capacity reserve for maintenance and disruptions	Moderate	10 to 15 min
80 - 100	**Highly impaired quality of service** with a low capacity reserve for maintenance and disruptions	High	15 to 25 min
> 100	**Infrastructure deemed to be saturated**	Saturation	> 30 min

This interpretation table summarizes our demonstration of the impact of the rate of utilization on the risk of delay and quality of service. We consider that once more than 70% of the theoretical capacity is used the quality of service is no longer guaranteed. It is thus not so much the density of traffic that determines the quality of service as the ability of the infrastructure manager to manage the network within its capacity reserves. One can imagine that optimized organization of maintenance and a more reliable operating system could considerably increase the level at which the quality of service comes under threat without increasing the risk of delay. Nevertheless, physical investment appears to be inevitable once the rate of utilization exceeds 80%. In this case everything depends on the dynamic of demand and how this is forecast to develop.

SECTION 5: EMPIRICAL APPLICATION

Five key variables have been identified for determining railway capacity in relation to demand: operation, management, infrastructure, rolling stock and pricing. Each variable is defined by parameters to analyse their impact on the rate of utilization. We propose to apply our model to an empirical study. For this we shall apply a simplified version of our method to the Paris-Lyon HSL. The traffic data were provided by the forward study presented in our introduction (LET-ENERDATA). According to the economic crisis, we have taken as our reference the lower projection to 2050. We have considered only mainline passenger traffic, which is estimated at 45 billion passenger kilometres on the Paris – Lyon HSL.

Table II – Supply hypotheses

	(reference)	**S 1 (rolling stock)**	**S 2 (operation)**	**S 3 (infrastructure)**	**S 4 (management)**	**S 5 (pricing)**
Passengers/train	450	**600**	450	450	450	450
Rate of multiple units*	1,3	**2**	1,3	1,3	1,3	1,3
Coefficient θ	0.8	**0.9**	0.8	0,8	0,8	**0,9**
Mileage in km	512	512	512	512	512	512
Hours of operation	6570	6570	6570	6570	**6935**	6570
Theoretical capacity	16	16	**20**	**32**	16	16
Coefficient k	0.75	0.75	**0,8**	0.75	**0.8**	0,75
Practical capacity	*12*	*12*	*16*	*24*	*13*	*12*
Concentration of traffic	1,5	1,5	1,5	1,5	1,5	**1,3**

*Specific to the French case where two TGV can be joined

According to this table, we define each variable by parameters. For rolling stock, we consider the number of passengers per train maximum that we balance by the load factor (coefficient θ) and the rate of multiple units. The load factor is also a good indicator to assess the strategy of operator for the service pricing and can be include in the pricing scenario. The hours of operation characterise the management scenario as the coefficient k. The theoretical capacity can be increased in the operating scenario with the coefficient k and in the infrastructure scenario. To finish, the concentration of traffic simulate the peak hour and is a good indicator in the case of the pricing scenario to assess the network manager strategy for infrastructure pricing.

The reference situation is based on the characteristics of the line in 2010. The mean theoretical capacity of a TGV is 450 passengers and the average load factor is 80%. The duration of operation in 2010 has been estimated at 18 hours per day (5.30 –23.30 hours). The theoretical capacity of the line is 16 slots per hour, i.e. one train every 4 minutes. In practice, the utilization rate is approximately 75% for a commercial capacity of 12 trains per hour. However, under exceptional circumstances this may increase to 80% but the risk of delay rises. This explains why we selected a value of 0.75 for the coefficient k. Last, we shall consider the peak period with a concentration coefficient of 1.5 comparing to the off peak period.

Compared with the reference scenario, in the “rolling stock” scenario (S2) the theoretical capacity in terms of the number of passengers doubles and the load factor increases, on condition that the operator decides to consolidate flows. The “operation” scenario (S3) considers the introduction of the new European rail signalling system (ERTMS) to increase the theoretical capacity and coefficient of resilience of a line by improving traffic management. The "infrastructure" scenario (S4) only considers the capacity in terms of the number of trains per hour. In this case, dualling the line corresponds to the project to build an HSL passing through the massif central[1]. The "management" scenario (S5) affects two parameters. Longer network operating hours mean that traffic could be spread of a longer period (19 out of 24 hours) and better management of traffic and the times set aside for works could improve the resilience of the network. Last, the pricing scenario includes the scarcity for operator and network manager. Operator maintains the same capacity per train but increase the load factor by a discriminant pricing. The network manager increases the infrastructure pricing by an internalisation of congestion cost to spread the peak hour. So this scenario consists to a maximisation of surplus user's by cost congestion.

Table II – Results

	Rate of utilization in 2050	**Capacity increase**	**Estimated cost**	**Time scale**
Référence	180%	-	-	-
S2 (rolling stock)	59%	+ 131%	Cycle of product life	Medium term
S3 (operation)	108%	+ 33%	€500 mln	Medium term
S4 (infrastructure)	68%	+ 100%	€12,9 bln	Long term
S5 (management)	160%	+ 13%	=	Short term
S6 (pricing)	104%	+ 30%	=	Short – medium term

These findings show the highly heterogeneous impact of the variables on the infrastructure utilization rate. If nothing changes between now and 2050, the utilization rate is due to increase 2-fold due to increasing demand. On this basis, if the amount of rolling stock is doubled, its load factor and management are optimized, it is possible to increase capacity by 131% which would help reduce the infrastructure utilization rate in 2050 by increasing returns. However, this scenario put highlights the challenge of comfort. On the other hand, changing the traffic management rules leads to a gain of 13% that can be interesting on short term. Dualling the infrastructure is an effective option in terms of capacity (a 100% increase). But it allows only constant returns for high cost. However, based on these results, no variable except infrastructure and rolling stock can provide a satisfactory utilization rate on the network. The future of the network must therefore be considered on the basis of a combination of these variables, so some choices will need to be made.

However, simply increasing capacity must not be the sole criterion. The stakeholders involved, the cost and the time-scale of actions are all factors that increase the complexity of the decision. In the case of rolling stock the rail operator makes most of the investment, in order to enlarge and improve its fleet. The manager is only involved at the margin unless the new rolling stock has a loading gauge that requires widening of the infrastructure (station, tunnel, sidings, etc.). On the contrary, if we consider the case of new infrastructure, it is the manager that is the main player with regard to managing the new services and financial involvement in the project. In this connection, as in most European countries the manager is

---

[1] The new Paris-Orléans-Clermont-Lyon line which is part of the Nation Transport Infrastructure Plan (Schéma National des Infrastructures de Transports (SNIT)).

under public supervision, the public authorities are directly involved in these decisions. Last, operation involves both the manager and the operators. For example, implementation of the new signalling system (ERTMS 2) requires the manager to pay for work on the tracks and the operators to pay for work on the locomotives.

Consequently, applying our method to the Paris-Lyon line, reveals two essential aspects of the study. First, each variable is relevant in a given domain. The choice must be made on the basis of the dynamic of demand as revealed by the analysis of the utilization rate that has been carried out previously. Second, our variables can also be considered as levers that are available to the stakeholders. Nevertheless, individually, none of the variables can provide a lasting solution. It is therefore necessary for the actions that involve them to be coordinated and coherent. This last point raises the issue of the governance of the railway system and the decisions made by its stakeholders.

CONCLUSION AND OUTLOOK

Our method proposes an evaluation of the utilization rate of railway infrastructure. The measurement of capacity is an essential part of such an evaluation. This concept has been widely explored in the academic literature. We have identified two types of research. The determinants of capacity were first of all analyzed and tested by engineers, as the railways were for a long time in the hands of vertically integrated monopolies. At the time, interest was focused on supply and the technical conditions that apply to it. The deregulation of railway networks that began in the United States (Staggeract, 1985) and continued in Europe (Sweden, 1989; Directive 91/440, 1991), opened up the sphere of railways to economists who rapidly began to investigate the issue of sharing capacity in terms of access to the network and the pricing of this public good which is subject to saturation. Analysis of the railway system thus made the transition from a supply-based approach to a demand-based approach.

Our method provides a demand-based approach to capacity. We have assumed that demand is completely satisfied, with the infrastructure changing in order to achieve this. Capacity is defined on the basis of five variables: the rolling stock, operation, management, infrastructure and pricing. The rolling stock expresses the operator’s decisions with regard to the management of demand, as the operator needs to find an equilibrium between the loading of trains and their frequency. Operation to a large degree determines infrastructure capacity. The theoretical capacity depends mainly on the principal characteristics of the infrastructure, but the capacity of the operating system also plays a role. However, the commercial capacity can vary according to decisions made by the infrastructure manager based on the management system and a desire to maximize revenue by means of a trade-off between the resilience of the system and the risk of delay. The infrastructure has specific characteristics in terms of the types and numbers of track which may to some extent determine the types of services (HSL or conventional line) and the general capacity of the system. Finally, pricing reveal strategy from stakeholders for capacity management. The operator can use scarcity to maximise the user’s surplus and network manager can use congestion pricing to maximise social utility of its network. Our method makes it possible to simulate separately how each of these variables affects infrastructure capacity according to demand.

The rate of utilization that results from the ratio between supply and demand is affected by two capacity thresholds. The first is the commercial capacity and is governed by the strategy implemented by the infrastructure manager. The second is the theoretical capacity of the infrastructure which is not entirely within the realm of the network manager. This threshold requires investment decisions on long term which should involve all the stakeholders in the railway system. The development of a new operating system frequently requires the operator to pay for modifications to the rolling stock. The same applies in the case of the construction of new infrastructure or the modification of existing infrastructure.

BIBLIOGRAPHY:

Abril, M., F. Barber, L. Ingolotti, M. A. Salido, P. Tormos and A. Lova (2008). An assessment of railway capacity. Transportation Research Part E, 44, 774-806.

Association of American Railroads (2007). National rail freight infrastructure capacity and investment study. AAR, Washington.

Burdett, R.L. and E. Kozan (2006). Techniques for absolute capacity determination in railways. Transportation Research Part B, 40, 616-632.

Carleson, L. (1957). A mathematical model for highway traffic. Nordisk Matematisk Tidskriff, 5, 176-180.

Crozet, Y. and I. Joly (2004). Budget Temps de Transport: les sociétés tertiaires confrontées à la gestion paradoxale du bien le plus rare. Les Cahiers Scientifiques du Transport, 45, 27-48.

D'Ariano, A., D. Pacciarelli and M. Pranzo (2008). Assessment of flexible timetables in real-time traffic management of a railway bottleneck. Transportation Research Part C, 16, 232-245.

Dingler, M., Y. C. Lai and C. Barkan (2009). Impact of train type heterogeneity on single-track railway capacity. Transportation Research Record, 2117, 41-49.

Dingler, M.H. (2010). Understanding the impact of operations and new technologies on railroad capacity. Ph.D. Dissertation, University of Illinois at Urbana-Champaign.

Dingler, M., A. Koenig, S. Sogin and C. Barkan (2010). Determining the causes of train delay. In: Proceedings of the 2010 Annual AREMA Conference, Orlando.

Ferreira, L. (1997). Rail track infrastructure ownership: investment and operational issues. Transportation, 24, 183-200.

Frank, H. (1963). Mathematical theories of traffic flow. Academic Press, New York.

Gibson, S., G. Cooper and B. Ball (2002). The evolution of capacity charges on the UK rail network. Journal of Transport Economics and Policy, 36, 341-354.

Gibson, S. (2003). Allocation of capacity in the rail industry. Utilities Policy, 11, 39-42.

Harker, P.T. and S. Hong (1994). Pricing of track time in railroad operations: An internal market approach. Transportation Research Part B, 28, 197-212.

Harrod, S. (2007). Rail capacity and management planning. Ph.D. Dissertation, University of Cincinnati.

Harrod, S. (2009). Capacity factors of a mixed speed railway network. Transportation Research Part E, 45, 830-841.

Higgins, A., E. Kozan and L. Ferreira (1995). Modelling delay risks associated with train schedules. Transportation Planning and Technology, 19, 89-108.

Higgins, A., E. Kozan and L. Ferreira (1996). Optimal scheduling of trains on a single line track. Transportation Research Part B, 30, 147-161.

Johnson, D. and C. Nash (2008). Charging for scarce rail capacity in Britain: A case study. Review of Network Economics, 7.

Kraft, E.R. (1982). Jam capacity of single track rail lines. Proceedings of the Transportation Research Forum, 23, 461-471.

Kraft, E.R (1987). A branch and bound procedure for optimal train dispatching. Journal of the Transportation Research Forum, 28, 263-276.

Kraft, E.R. (1988). Analytical models for rail line capacity analysis. Journal of Transportation Research Forum, 29, 153-162.

Krueger, H. (1999). Parametric modeling in rail capacity planning. In: Proceedings of the 1999 Winter Simulation Conference, Washington.

Lach, D. C. and G. R. Skelton (1968). SIMTRAC – A simulation of train dispatching on single-track railway. In: CORS National Conference, Toronto.

Lai, Y.C. (2008). Increasing railway efficiency and capacity through improved operations, control and planning. Ph.D. Dissertation, University of Illinois at Urbana-Champaign.

Lai, Y.C. and C. Barkan (2009). An enhanced parametric railway capacity evaluation tool (RCET). Transportation Research Record, 2117, 33-40.

Landex, A. and A. H. Kaas (2005). Planning the most suitable travel speed for high frequency railway lines. In: the 1st International Seminar on Railway Operations Modelling and Analysis, Netherlands.

Landex, A., A. H. Kaas, B. Schittenhelm and J. Schneider-Tilli (2006). Pratical use of the UIC 406 capacity leaflet by including timetable tools in the investigations. Computers in railways, X, 643-652.

Liotta, G., P. Cicini and P. De Luca (2009). A simulation-based approach for estimating the commercial capacity of railways. In: Proceedings of the 2009 Winter Simulation Conference, Washington.

Marlot, G. (2012). Estimation du coût marginal de la congestion ferroviaire. Working Paper, Paris.

Martland, C. D. and G. Hutt (2005). Analysis of potential approaches to interline capacity flow management, RaiLinc, Cary.

Nash, C. A. and B. Matthews (2005). Rail infrastructure charges – the issue of scarcity. Working paper, University of Leeds.

Nilsson, J. E. (1999). Allocation of track capacity: Experimental evidence on the use of priority auctioning in the railway industry. International Journal of Industrial Organization, 17, 1139-1162.

Nilsson, J. E. (2002). Towards a welfare enhancing process to manage railway infrastructure access. Transportation Research Part A, 36, 419-436.

Petersen, E. R. (1974). Over the road transit time for a single track railway. Transportation Science, 8, 65-74.

Prokopy, J. and R. Rubin (1975). Parametric analysis of railway line capacity. Technical report U.S. Department of Transportation, Federal Railroad Administration.

Quinet, E. (2003). Short term adjustments in rail activity: the limited role of infrastructure charges. Transport Policy, 10, 73-79.

Schafer, A. and D. G. Victor (2000). The future mobility of the world populations. Transportation Research Part A, 34, 171-205.

Schafer, A. (2001). Long-Term Trends in Global Passenger Mobility. National Academy of Engineering Website, www.nae.edu/.

Schlake, B., C. Barkan and R. Edwards (2011). Train delay and economic impact of railroad rolling stock in-service failures. In: Transportation Research Board 90th Annual Meeting, Washington.

SETRA (2009). Le transport ferroviaire de voyageurs sur le réseau ferré national français. Synthèse des connaissances, Bagneux.

DGITM (2011). Schéma national des infrastructures de transport. Avant-projet consolidé, Paris.

UIC (2004). Capacity: Leaflet 406. In: 1st edition International Union of Railways, France.