Towards a rail congestion formalization from a consumer perspective

Maria PEREZ HERRERO
Junior Scholar
LET, Université de Lyon 2- SNCF Réseau

1 Introduction

Rail capacity constraints is an increasing important issue for the rail infrastructure managers. Demand for rail transport grows in metropolitan areas and the quality service, in terms of regularity, become a central issue for the densely populated areas. Otherwise, the expansion of rail capacity faces a range of obstacles and financial challenges. Finding an equilibrium between rail services supply and quality service, considering the available resources, is one of the major challenges for the rail infrastructure managers.

Line capacity is essentially what the infrastructure managers have to sell as their final service. However, the definition of rail capacity constraints 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).

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 positive related to the density of traffic (congestion).

The measure and the value of rail capacity constraints should play a key role in all the decisions related to capacity investments. Nonetheless, the existent analyses does not allow to consider the value of capacity constraints in the rail infrastructure manager’s decision-making process.

The aim of this paper is to provide a few hints for formalizing the economic fundamentals of rail congestion from a consumer perception. The structure of the paper is as follows. Section 2 details precisely the real issues at stake when facing problems of capacity constraint in rail transportation, describing the actual timetabling process in France. Section 3 describes a consumer model taking into account all the trade-off concerning capacity constraints for consumers. The model identify the consequences for users of having different rail frequencies in terms of of service provided (Mohring effect
on the scheduling cost) and delay cost. Section 4, identifies what is the analytical optimal frequency (per unit of time) which maximises passengers utility considering the previous model. Section 5 provides some numerical examples that illustrate the relationships between the optimal frequency and the other parameters. In a further research, this section should be developed and completed with a calibration model. The aim of the calibration model would be to determine a feasible interval of optimal frequencies, considering the variability of the others parameters.

2 Current timetabling process: the French experience

Rail timetabling process is different between countries. Each national rail system has its own structure and its own methods in timetabling process.

In the timetabling process design, the rail infrastructure manager embeds the objective of having reliable train paths. From the infrastructure manager perspective, this objective is integrated in the travel time path or in the time buffer between two trains. The first one allow to insure the robustness of an individual train path, and the second one, the robustness of the diagram (Verchere et Djellab, 2013)

In this section, let us detail the actual scheduling process in France

2.1 Train path robustness and recovery margin

Train path travel time (IN 1514, ex Règlement S2C, article 101) is determined considering technical and commercial characteristics. Following the internal document “Détermination et confection des horaires” (DCF-DPS Supervision et Support. Réseau Ferré de France, 2006). Train path travel time is the sum of four parameters:

- Basic running time
- Standing time
- Recovery margin
- Extra time needed to enable compilation of the diagram (stops, extensions to requested standing time, works possessions and/or other train paths, etc.).

Train path travel time is so defined as:

1. The basic running time is the result of calculations for the train indicated by the applicant (traction and rolling resistance characteristics) on the infrastructure selected for the calculation by the timetable. This is a parameter that is designed depending on various operational aspects. The first one is the track configuration. The network can consist of single, double, triple or even more tracks, 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.

2. Standing times for commercial or technical stops asked by the train operator companies.

3. The recovery margin is extra time (added to the basic running time) to allow for:
   - slippage and production or traffic contingencies (IM allowance and applicant’s allowance) (around 2 min/100 km on the ordinary lines);
   - time that may be lost in the event of a normal maximum number of speed restrictions for works, or the enforcement of IPCS counterflow working during maintenance operations (around 2,5 min/100 km on the ordinary lines).

The recovery margin will be spread over the whole route in such a way as to establish a compromise between a part of the time that will be actually spread in this way and another, less specific part that will be kept for the end of the journey.

The normal recovery margin is 5 movements on ordinary lines, other than in the specific cases described in the line reference documents. Special margins may be applied by RFF, on its own initiative or at the request of a stakeholder, to certain lines, in particular:
   - limited to 3 min/100 km for certain designated passenger trains;
   - increased to 5.5 min/100 km for train paths with a maximum speed of less than or equal to 100 km/h, or during certain maintenance possession periods announced by the infrastructure manager.

4. In some cases, extra minutes, also to be added to the basic running time in the event of temporary speed restrictions, will be announced in advance of the annual timetable compilation period.

Railway traffic timetables contain margins to make them robust, and enable trains to recover from certain delays. How effective these margins are, depends on their size and location as well as the frequency and magnitude of the disturbances that occur.

The recovery margins added to the basic running time allow to propose a reliable train service with robust travel times subjected to low variations. In return, however, the train travel time is increased systematically in the timetable process. The aim of this analysis is not to determine if the volume of the actual recovery margins are optimal, but to emphasize that the shadow costs for a more reliable train services are longer travel times.

2.2 Train diagram robustness

In the precedent section, we have presented the practices that allow to insure robustness for an individual train path. However, train paths must be compatible between
them an they are plotted so that two trains running on time on “standard” infrastructure do not come into conflict. First of all, there exists a minimum time interval between two trains, which depends on the signal system. In order to insure robustness on the train diagram, it is possible to add some extra time between trains:

- Un extra time (in general de 35 seconds) is added for insuring that train drivers always see a green signal. This time is designed by the Greek letter $\chi$.
- Rounding results of the minimal time interval calculation.
- Time buffers between planned train paths with the purpose of assure timetable robustness. Buffer times reduce the risk of transferring delays between trains or they reduce the size of the knock-on delay transferred from one train to the following train. There are essentially two ways to fix time buffers. The first one consists of adding an extra interval to the minimal time interval. For example, in a HSL it is possible to run a train every four minutes; however, in some line this interval increases to five minutes in order to add an extra time for recovering from risk. The second way for increasing robustness is by adding an empty slot between some trains. For example, after three slots with HS trains, the fourth one is empty.

The purpose of these kinds of practices is to ensure the train robustness (they minimize the probability for the passengers to be delayed). However, they reduce capacity that could be potentially expended for running extra trains. Buffer times in railway timetables are a trade-off between capacity consumption and punctuality levels.
This description permits to understand how margins in the scheduling process influence total travel time for each slot or/and the line capacity. In fact, looking for a robust timetable would have an impact on scheduling travel time or/and in capacity. In Figure 1, the left-hand side reflects the recovery margin practice. Since this is applied to all configurations, it is thus assumed in the central part and in the right-hand side (all slopes coincide with the one defined by the recovery margin parameter). The right-hand side illustrates the empty slot method, while the green line in the medium diagram exemplify the extra interval practice.

3 The theoretical model

The methods described in the precedent section demonstrate that the actual train timetable process in France considers a quality logic for the service supplied. Ceteris paribus, there exists a trade-off between train path robustness and travel time or between train diagram robustness and capacity. Nevertheless, the level of these trade-off are not fixed today.

The goal of this paper is to establish a simplified model that reflects all the trade-off
concerning capacity constraints for consumers. The model aim to identify what is the optimal frequency (per unit of time) for passengers bearing in mind all the adjustment described.

We consider a simplified network with a double track line with homogeneous traffic between two train stations $A$ and $B$.

A benevolent rail infrastructure manager searches to maximize utility for passengers.

In this section, we target to establish the net utility function for passengers. We will consider that passengers experiment a gross utility $\sigma$ from traveling by train between cities $A$ and $B$ but they incur also some costs when they decide to make a trip.

For the general specification of the model, a number of assumptions are made, that will now be presented.

- **Assumption 1.** As rail transport is a scheduled transport mode, the infrastructure manager establishes a frequency $f$ (number of trains/unit time) between cities $A$ and $B$.

- **Assumption 2.** Demand $N$ is uniformly distributed throughout the unit time $T$. We consider $N = kT$.

- **Assumption 3.** Travel time represents a cost for the passenger.

  As detailed before, train operator companies, introduce some recovery time $\zeta$ to control uncertainty. In other words, they announce longer travel journeys than the basic travel time. The scheduled travel time for a passenger is equal to the basic travel time, $J$, plus the train recovery time $\zeta$.

  The cost associated to the scheduled travel time is given by Eq.(1), where $vot$ is the value of time for the scheduled time.

  \[
  \text{Scheduled travel time cost} = vot(J + \zeta)
  \]  

- **Assumption 4.** Trains are equally spaced during the considered unit of time. The difference between the preferred travel time by users and the scheduled travel time fixed by the infrastructure manager, represents a scheduling cost for users.

  Following the notion on the location models of Hotelling (1929) et Salop (1979), we consider that each consumer has a most preferred travel time $\bar{a}_t$. In a transport mode where frequencies $f$ are discrete and fixed in advance by the infrastructure manager, passenger must adjust their most preferred travel time to the scheduled travel times. This difference between times generates a disutility on each passenger. Imagine that a passenger most preferred travel time is $\bar{a}_t$. We assume that the first train arrives at 0, and the last train at $T$. The value $T$ is the operating time interval.
As there is not a train which arrives exactly at time $\bar{a}_t$, the passenger would choose between taking the train arriving before - being then ahead of time at his destination - or taking the train arriving after - being then late.

According to the road congestion literature, Vickrey (1969), Arnott, De Palma, and Lindsey (1990), we consider that arriving at a different time than the preferred one represents a cost for travellers. However, the phenomenon is different because taking a train is not random. This contrasts with travel by auto-mobile, which can be initiated any time. On auto trips, passengers make a trade-off between travel time (trying to avoid peak-period congestion) and schedule delay for trip timing decisions. However, on scheduled transports, most transit users suffer a schedule delay even if the transit system is reliable and adheres perfectly to the timetable (de Palma and Lindsey, 2001).

In this paper, we delimit the cost due to this time imbalance as the “scheduling cost”, with the following definition:

- If a passenger decides to arrive at $T_1$, he/she should leave early his others activities (wake up early, leave work early, etc). Moreover, he/she will arrive before that he would like, so he/she should wait at the station. The cost is $v_o \alpha_1 t$, (where $v_o$ is the value of time for the scheduled time and $\alpha_1$ the schedule delay multiplier of arriving early (before $\bar{a}_t$))

- If he/she decides to arrive at $T_2$, he/she could wake up or stay at home calmly, but he/she will arrive later than preferred at the destination, with a cost $v_o \beta_1 (r-t)$, (where $\beta_1$ is the schedule delay multiplier of arriving late (after $\bar{a}_t$))
In order to calculate the disutility created, we must first estimate the location $a_i^*$ of the passenger indifference between both alternatives. The Figure 4 represents the utility function as a function of the preferred arrival time. This utility function is always less than or equal to 0. Of course it is 0 if the preferred arrival time coincides with the actual arrival of a train.

$$a_i^* = \frac{r\beta_1}{\alpha_1 + \beta_1}$$

Once determined the time $a_i^*$, a passenger compares its preferred arrival time $\bar{a}_i$ with $a_i^*$ and decides which train to choose. If $\bar{a}_i < a_i^*$ he/she would choose the previous train. Otherwise, he/she would choose the next train (with a utility $U(\bar{a}_i)$, being $U$ the function in the picture). Given his/her decision, he/she assumes the disutility associated to the difference between $\bar{a}_i$ and the chosen train arrival. As shown before, it equals the slope of the straight lines defined by equations (2) and (3) respectively. As we have considered that the demand is uniformly distributed throughout the interval $T$ (Assumption 2), demand for the interval $r$ equals $kr$. The total disutility for passengers in the interval $r$ equals the total area of the triangle $^1$.

$$\text{Base} = rk = \frac{T}{f} * k$$

$$\text{Height} = \frac{r\alpha_1\beta_1}{\beta_1 + \alpha_1}$$

Scheduling cost (total users interval $r$) = $$\left[ \left( \frac{T}{f - 1} \right)^2 \frac{k\alpha_1\beta_1}{2(\beta_1 + \alpha_1)} \right] v o t$$

In consequence, the average consumer’s cost for time adjustment can be represented by a function of the form:

---

1. It is assumed that individuals are identical except for their desired travel times.
Average scheduling cost = \[
\left\lfloor \frac{T}{f - \frac{1}{2} \beta_1 + \alpha_1} \right\rfloor \text{vot}
\] (8)

The equation 8 shows that the scheduling cost diminish with frequency \( f \). This relationship underlines an identified effect for transports where timetabled are scheduled in advance, known as the “Mohring effect”.

The relationship between frequency and waiting time costs has its origins in an analysis realized for public transport, and particularly for buses Mohring (1972). In the context of public transports, users are considered to arrive randomly. Their waiting time cost at the bus stop depends on the bus service frequency. A higher demand level in an given geographical area will generate a diminution in the total travel cost for users, due to the increment on frequencies.

In rail transport, timetable is fixed in advance and users are not supposed to arrive randomly, so a variation on frequency does not mean a variation on waiting time cost at the station. By contrast, a change on frequencies and on timetable, involve a variation on the scheduling cost. The Mohring effect, in a context of transport scheduled in advance, can be interpreted as a variation on the scheduling cost when the frequencies change.

In assumption 4, it has been considered that users experiment a scheduling cost when they decide to travel due to the impossibility of adjusting perfectly their preferred travel time. This delay is independent on travel time reliability, which will be considered in the next assumptions.

- **Assumption 5.** In some cases, stochastic delays can increase the scheduled travel time. Delay propagation depends on traffic density.

As detailed before, the infrastructure manager internalizes some expected delays when it designs train diagram, but this does not mean the absence of delays in the network: some stochastic delays exist.

To be able to relate reliability of trains to capacity utilization, useful indicators of reliability are needed. Carey (1999) presents an insightful analysis of the mechanism behind delays. He considers two types of delays: exogenous or primary delays and knock-on or secondary delays. Exogenous delays are due to events such as breakdown or failure of equipment or infrastructure, delays in passenger boarding, lateness of operations or crews, etc. Exogenous delays are generally not caused by the schedule. In contrast, knock-on delays are due to exogenous delays and the interdependence in the schedule. Under high utilization, one delayed train can cause delays to several other trains over a large area and a long period of time. Knock-on or secondary delays can be reduced by scheduling, for example by giving more headway to trains which are prone to exogenous delay.

Following Villemeur, Billette, Ivaldi, Quinet, and Urdanoz (2015), we consider that stochastic delays \( \varepsilon \) are \( \varepsilon > 0 \). Distributions of delays are considered to be (exogenously) given.
As Villemeur and al. (2015), we do not consider that primary delays depend upon the pattern of flows. Our intuition is that the probability of a primary delay is given and independent on the number of flows (technical problems or human errors are not function of the number of trains running in our link). The recovery time \( \zeta \) considered in the scheduling process can allow for recuperating from an incident in some cases.

Nevertheless, in contrast to the Villemeur and al. (2015), we consider that the specification of the model must also reflect congestion issues. In fact, the origin and the probability of an incident are effectively independent on the number of flows, but the consequences of these events are strongly linked to the number of flows (trains/unit time). When a train track is highly used, an additional slot increases the consequences of delays, due to a reduction in the ability to recover from an incident. It means that when traffic is high, the probability of spreading delays is higher and in consequence, the total effects of delay are larger. As in airports, rail congestion exhibits a cascade-type effect: one single delay may generate an impact which accumulates over the next trains.

In order to control for delay propagation, a buffer time \( \gamma_b \) between trains is introduced in the scheduling process. High capacity consumptions results in higher risks of consecutive delays. If there is enough buffer times between two trains, small delays will not affect the successive train(s). When a primary delay propagates to another train, a secondary delay can arise. Following Landex and Nielsen, (2008), the propagation delay description in this model assumes a double track line with homogeneous one way operation on each track (meaning that both the speed and the buffer time are constant).

The size of the buffer time between two train can be expressed as the difference between \( \gamma_h \) which is the effective interval time and \( \gamma_{h,\text{min}} \) which is the minimum time interval.

\[
\gamma_b = \gamma_h - \gamma_{h,\text{min}} \tag{9}
\]

Following the notation, we can write the maximal capacity/frequency of the line as:

\[
f_{\text{max}} = \frac{T}{\gamma_{h,\text{min}}} \tag{10}
\]

And the given frequency \( f \):

\[
f = \frac{T}{\gamma_h} = \frac{T}{\gamma_{h,\text{min}} + \gamma_b} \tag{11}
\]

The delay function considered in this paper aims to combine the two precedent approaches.

2. Reminder: We consider un homogeneous and uniformly distributed traffic
Let’s consider a stochastic delay $\varepsilon$, independent on traffic flows. If $\varepsilon < \zeta$, the recovery time can recover the incident, and passengers do not suffer a delay. However, if $\varepsilon > \zeta$, recovery time would not be enough to recover the time lost from the incident, and passengers would suffer a delay ($d_{1,i}$). The amount of delay for the first train is:

$$d_{1,i} = \varepsilon - \zeta$$  \hfill (12)

Moreover, if $d_{1,i} > 0$, the primary delay can be propagated to next trains, depending on the buffer time level between trains. The amount of delay propagation, or consecutive delay for the following train $d_{2,c}$, can be calculated as:

$$d_{2,c} = \begin{cases} d_{1,i} - \gamma_b & \text{si } \gamma_b < d_{1,i} \\ 0 & \text{sinon} \end{cases} \hfill (13)$$

$$d_{2,c} = \left\{ \begin{array}{ll} d_{1,i} - \gamma_b & \text{si } \gamma_b < d_{1,i} \\ 0 & \text{sinon} \end{array} \right. \hfill (14)$$

If the buffer time $\gamma_b$ is larger than or equal to the delay $d_{1,i}$, the delay will not lead to a consecutive delay of the successive train, $d_{2,c}$ will then be less than or equal to zero. Formula (13) can be generalized to calculate the consecutive delay for any of the following trains where there are no more initial delays.

$$d_{j+1,c} = d_{1,i} - n\gamma_b$$  \hfill (15)

In formula 15, $n$ is the number of trains receiving consecutive delays. By setting the consecutive delay $d_{n+1,c}$ equal to zero (meaning that the last train will receive no consecutive delay), it is possible to calculate the number of trains needed before the trains again run on time:

$$n = \frac{d_{1,i}}{\gamma_b}$$  \hfill (16)

As a train is either delayed or on time, the decimal numbers in precedent formula should be truncated.

$$n = \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor$$  \hfill (17)

Knowing the number of trains $n$ receiving consecutive delays, it is possible to calculate the total delay, equal to the sum of consecutive delays and the initial delay:

$$\sum d = d_{1,i} + d_{2,c} + d_{3,c} + d_{4,c} + \cdots + d_{n+1,c} = d_{1,i} + \sum_{k=1}^{n+1} d_{k,c}$$  \hfill (18)

Combining formula 15 and 18, the total delay can be rewritten as:

11
\[ \sum d = d_{1,i} + d_{1,i} - \gamma_b + d_{1,i} - 2\gamma_b + \cdots + d_{1,i} - n\gamma_b = (n+1)d_{1,i} - \frac{n}{2}(n+1)\gamma_b \] (19)

Combining formula 17 and 19, the total delay can be calculated based on the initial delay \( (d_{1,i}) \) and the buffer time \( (\gamma_b) \):

\[ \sum d = \left( \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor + 1 \right) d_{1,i} - \frac{1}{2} \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor \left( \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor + 1 \right) \gamma_b \] (20)

This formula is based on the propagation delay function detailed in Landex and Nielsen,(2008) and has been completed considering that recovery time \( (\zeta) \) has an influence on the primary delay function. It reflects that the recovery time has an impact on the initial delay and in consequence on the propagation of delay. In our sense, this approach gives a comprehensive notion and completes the functions provided by Villemeur al.(2015) and Landex and Nielsen,(2008) .

The total delay function 20 reveals that adding a buffer time \( (\gamma_b) \) diminishes the total delay. In return, that limits total capacity, so the frequency supplied.

Delays increase par definition travel time for users (Assumption 3). Delays as unexpected events, present higher costs for passengers than costs related to scheduled travel times. We consider that there exists a lateness penalty multiplier \( \beta_2 \) associated to the \( vot \).

Taking into account that we can rewrite 20 as the total delay cost for passengers as :

\[ D = vot\beta_2 \left( \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor + 1 \right) d_{1,i} - \frac{1}{2} \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor \left( \left\lfloor \frac{d_{1,i}}{\gamma_b} \right\rfloor + 1 \right) \gamma_b \] (21)

Combining function 11 and 21, if it is feasible to express the total delay for the whole of trains as a function of the frequency \( f \):

\[ D = vot\beta_2 \left( \left\lfloor \frac{d_{1,i}}{\frac{T}{f} - \gamma_{h,min}} \right\rfloor + 1 \right) d_{1,i} - \frac{1}{2} \left\lfloor \frac{d_{1,i}}{\frac{T}{f} - \gamma_{h,min}} \right\rfloor \left( \left\lfloor \frac{d_{1,i}}{\frac{T}{f} - \gamma_{h,min}} \right\rfloor + 1 \right) \left( \frac{T}{f} - \gamma_{h,min} \right) \] (22)

Considering function 22, we can precise the average delay for a passenger as :\(^3\):

\(^3\) The average delay for a consumer is equivalent to the average delay for a train.
The passenger optimization problem: analytical solution

Considering the assumptions described in the precedent section, it is possible to define the optimal frequency that maximizes the net consumer utility:

\[
\text{Max}_f \ U = \sigma - \left[ \frac{T}{f - 1} \frac{\alpha_1 \beta_1}{2(\beta_1 + \alpha_1)} \right] \text{vot} - \text{vot}(J + \zeta)
\]

\[
- \frac{1}{f} \left[ \text{vot} \beta_2 \left( \left( \frac{d_{1,i}}{T - \gamma_{h,min}} \right) + 1 \right) d_{1,i} - \frac{1}{2} \left( \frac{d_{1,i}}{T - \gamma_{h,min}} \right) \left( \left( \frac{d_{1,i}}{T - \gamma_{h,min}} \right) + 1 \right) \left( \frac{T}{f} - \gamma_{h,min} \right) \right]
\]

(24)

The benevolent infrastructure manager wish to maximize the net utility for passengers from travelling by train between the cities pair A and B considering the associated costs: scheduling costs, travel time costs and the delay costs. The infrastructure manager looks for an optimal frequency, knowing that, ceteris paribus, high frequencies means less scheduling costs (second right hand term), but correspondingly, more expected delays (last right hand term).

The first-order necessary conditions are:

\[
\frac{\partial U}{\partial f} = 0 = \beta_2 \frac{d_{1,i} \text{vot}}{2 f^2} \left( \frac{d_{1,i}}{T - \gamma_{h,min}} + 1 \right) - \frac{\beta_2 \frac{d_{1,i}^2 \text{vot}}{T - \gamma_{h,min}}^2}{2 f^3} + \frac{\alpha_1 \beta_1 (\beta_1 + \alpha_1) T \text{vot}}{2 f^2}
\]

(25)

Each additional \( f \) diminish the marginal scheduling cost (last right hand term) and increases at the same time the marginal delay cost (first two right hand terms).

At the equilibrium, the infrastructure manager would chose \( f^* \) optimal from the consumer perspective. This frequency establish that the marginal cost of scheduling adjustments, equals the marginal cost of delays.
\[ f^* = -\frac{d_1 \sqrt{\left( \alpha_1 \gamma_{h,\text{min}}^2 \beta_1^2 + \alpha_1^2 \gamma_{h,\text{min}}^2 \beta_1 \right) \beta_2 T + \gamma_{h,\text{min}}^2 \beta_2^2 d_{1,i} T}}{\left( \alpha_1 \gamma_{h,\text{min}}^2 \beta_1^2 + \alpha_1^2 \gamma_{h,\text{min}}^2 \beta_1 \right) T - \gamma_{h,\text{min}} \beta_2 d_{1,i}^2 + \gamma_{h,\text{min}}^2 \beta_2^2 d_{1,i} T} \]

\[ \gamma_b^* = \frac{T}{f^*} - \gamma_{h,\text{min}} \]  

5 The passenger optimization problem: some numerical examples

This section will present some numerical results that further illustrate the properties of the model. In order to better understand the relationship between the optimal frequency and the other parameters of the model, the following figures illustrate how varying their numerical values affects the optimal frequency.

![Figure 5](image-url)

**Figure 5**

As we observe in figure 5 and figure 6, scheduling costs have a positive relationship with optimal frequency for passengers. If the cost of not having frequencies at the desired
travel times were high for passengers, they would like to have higher frequencies for their travel.

\[\text{Figure 6 –}\]

In contrast, in figure 7 we remark a negative relationship between the lateness penalty and the optimal frequency. If the penalty of arriving late, after the scheduled arrival time, were elevated, passengers would prefer to have lower frequencies. In fact, passengers know that if frequencies are high, the probability of been late is higher too.

\[\text{Figure 7 –}\]
Figure 8 illustrates a negative relationship between the initial delay and the optimal frequency for passengers. If the initial delay is important, it would be more difficult for the network to recover and the delay propagation would be easier. In this case, we have considered that delay is known in advance for passengers. In a further version, it will be interesting to model a probabilistic delay.

To finish, figure 9 shows a negative relationship between the minimum headway between two trains and the optimal frequency. Intuitively, if the headway between two trains is higher, the number of trains in one unit of time would be lower.
6 Conclusion

Rail capacity constraints analysis it is a contemporary challenge for the rail infrastructure managers. The description of the current timetabling process point out that, all things being equal, there exists a relationship between the traffic volume, delays and train path scarcity. From a consumer perspective, rail capacity constraint can be expressed in two complementaries ways, but not exclusives. On the one hand, there exits “congestion” (relationship between traffic and delays) and on the other hand, “scheduling cost”( the impossibility to travel at the preferred travel time.

The microeconomic model proposed in this paper demonstrate that there exists implicitly a trade-off between the supplied frequency (and in consequence the diminution of the scheduling costs) and delays in the current timetabling process. However, the link between these two variables has not been formalised and measured until present. The aim of this formalisation is to define the frequencies $f^*$ which maximizes the consumer surplus, considering all the relationships specified.

Optimal frequency depends on several parameters and their analysis allows to illustrate the properties of the model. The paper proposes a first step on the analysis of these relationships, but the results should be developed in further researches. The results of the theoretical model could be evaluated from an empirical point of view using a calibration model. The aim of the calibration model is to determine a feasible interval of optimal frequencies, considering the variability of the others parameters. The calibration of the parameters will be based on the empirical values available on the academic literature (scheduling costs, delay cost multiplier and values of time) and on the infrastructure manager practices described before. This further analyses could allow to identify the
value of frequencies and buffer times, that under certain values of the other parameters, maximizes the consumer surplus.

7 Bibliography