Ride into the danger zone: avoiding the wrong frequency for an express bus service.

Homero Larrain · Juan Carlos Muñoz

Abstract Express bus services provide a direct ride for passengers travelling between two distant points over a transit corridor. In this work we model an ideal transit corridor where one regular service operates in tandem with an express service. We model the problem were a city planner wants to determine the optimal frequencies to offer for each service, taking into account that passengers will choose how to travel rationally and that buses have a limited capacity. Some mathematical characteristics of this problem are explored, including an analysis of its feasible region. This led us to observe the existence of a specific range of frequencies for the express service (i.e., a “danger zone”) where queues appear in the system, bringing down its level of service.

Keywords: Public transport · Express services · Frequency optimization

1 Introduction

Express bus services are services that skip some of the stops along their routes to provide a faster ride on particularly demanded trips on a corridor. The problem of identifying these services and setting their frequencies has caught a lot of attention lately, due to its relevance in the planning of an efficient bus corridor in the context of the now ubiquitous paradigm of Bus Rapid Transit. Some works report case studies on the effects of implementing this type of services; see, for instance, Ercolano (1984), Silverman (1998), Tétreault and El-Geneidy (2010) and El-Geneidy and Surprentant-Legault (2010). Many authors propose different formulations and algorithms for solving the express service design problem, like Jordan and Turnquist (1979), Furth (1986), Leiva el al. (2010), Sun et al. (2008), Chen et al. (2012), Chiraphadhanakul and Barnhart (2013), Larraín et al. (2015) or Soto et al. (2017). These models, although useful for solving particular instances of the problem, rely on very complex formulations that do not allow for an in depth analysis of the problem from a mathematical standpoint.
This work proposes a simple scenario that lends itself for different analyses. In particular, we prove that, against what could be expected, increasing the frequency of an express service can bring down the level of service of the users of this service, and trigger queues that were previously inexistent. This could be interpreted as a phenomenon similar to the Braess’ paradox, in which increasing the capacity of a transport network increases the average travel cost of the users. In our case it will be shown that the described phenomenon occurs only in a specific range of frequency values, which we call the “danger zone” of express services.

2 Definition of the problem

Consider one direction of a public transport corridor over which two different bus services operate. The first one is a regular, “all-stop” service, which as its name implies stops at every bus stop along its route. The second service is an express service, which connects the two extreme stops of the corridor, skipping all the other stops and thus saving \( \tau \) units of travel time.

The demand for trips on this corridor at a given period of time can be classified into two groups. First, there are some users travelling from the first stop of the corridor to the last one, and that can thus use the express service for a direct ride. These users may still take the all-stop service if it shows up first, depending on the time savings involved. Then, there is the rest of the trips on this direction. Assuming for the simplicity of this analysis that users are not willing to transfer, these trips will only use the all-stop service. We will denote the captive demand of all-stop services as \( T_A \), and the demand that can opt to use the express service as \( T_E \). Also for simplicity, we will assume that the demand in the opposite direction of the corridor is negligible.

Captive users from the all-stop service will board and alight at different stops of the corridor, defining a base load profile for this service. We will denote the critical load of this profile, i.e., the number of trips on the most demanded segment of the all-stop-only profile, as \( L_A \). Similarly, we define \( L_E \) as the load on every arc from trips in \( T_E \). If a strict capacity \( b \) of the buses is assumed (same for every bus in every service), then \( P_A := L_A / b \) is a lower bound for the frequency needed in the all-stop service during the planning horizon to meet this demand. In the case of the express service, \( P_E := L_E / b \) is the minimum frequency that can serve all the trips going from the first to the last stop of the corridor.

Users that can take the express service can choose between two strategies to perform their trip. The first one is to wait exclusively for the express service and take it. If the frequency of the express service is \( f_E \), the expected travel time for this strategy can be modeled as:
\[ C_i = t_E + \frac{\theta}{f_E} \]  

In this expression, \( t_E \) represents the travel time of the express service. The term \( \theta/f_E \) models the expected waiting time, assuming that buses arrive to the stop following a Poisson process, while passengers have no information about bus arrival times so both processes are independent.

The second strategy would be to take the first bus to arrive to the stop. In this case, the expected travel time can be estimated as:

\[ C_{II} = \frac{f_A t_A + f_E t_E + \theta}{f_A + f_E} \]  

This expression is the sum of the expected travel time and the expected waiting time, considering that the probability of using a particular service is proportional to its frequency. If users minimize their expected travel time, it can be proved that they will wait exclusively for the express service if and only if the frequency of this service is higher than a critical frequency \( \hat{f}_E \), defined as \( \hat{f}_E = \theta/\tau \), where \( \tau \) corresponds to the travel time saving of taking the express service over the all-stop.

### 3 Characterization of the feasible region of the problem

Using \( \tau \), for any given values of \( f_A \) and \( f_E \) we can now predict the utilization of each service (assuming that bus capacity never prevents a user from taking a bus), and thus decide if the capacity of the services is enough to carry their demand. This yields a feasible region for \( f_A \) and \( f_E \) that we will denote as \( \Omega \). In order to characterize \( \Omega \), we will divide it in two regions: \( \Omega_I \), where \( f_E \leq \hat{f}_E \), and \( \Omega_{II} \), where \( f_E > \hat{f}_E \).

In region I, where \( f_E \leq \hat{f}_E \), users in \( T_E \) will take the first service to show up. This means that the all-stop service has to have enough capacity to carry \( L_A \) plus the part of \( L_E \) that will take this service. This condition can be expressed as:

\[ f_A \geq P_A + \frac{f_A}{f_A + f_E} \cdot P_E \Leftrightarrow f_E \geq \frac{f_A}{f_A - P_A} P_E - f_A \]  

A similar capacity constraint can be stated for the express service:

\[ f_E \geq \frac{f_E}{f_A + f_E} \cdot P_E \Leftrightarrow f_A + f_E \geq P_E \]  

Given that \( f_A/(f_A - P_A) > 1 \), condition (2) is redundant with (1). Hence, \( \Omega_I \), can be defined by the intersection of regions (1) and (3):
\[0 \leq f_E \leq \hat{f}_E\] (3)

For region II, where \(f_E > \hat{f}_E\), every user in \(T_E\) will take the express service. Therefore, bus capacity constraints can be stated as follows:

\[f_A \geq P_A\] (4)
\[f_E \geq P_E\] (5)

Expression (5) can be replaced by (6):

\[f_E \geq \max\{P_E; \hat{f}_E\}\] (6)

Thus, region \(\Omega_{II}\) can be defined as the intersection of regions (4) and (6).

Note that both \(\Omega_i\) and \(\Omega_{II}\) and hence \(\Omega\) can be determined by three frequencies: \(P_A\), \(P_E\) and \(\hat{f}_E\). Regions \(\Omega_i\) and \(\Omega_{II}\) will be contiguous when \(P_E \leq \hat{f}_E\), and disjoint when \(P_E > \hat{f}_E\). Figure 1 illustrates \(\Omega\) for the case where \(P_E > \hat{f}_E\).

Figure 1 shows that when \(P_E > \hat{f}_E\) an infeasibility zone appears between \(\Omega_i\) and \(\Omega_{II}\), defined by \(\hat{f}_E < f_E < P_E\). In this zone all the passengers in \(T_E\) will want to take the express service (because \(\hat{f}_E < f_E\)), but the frequency of the express service will not be enough to meet its demand (\(f_E < P_E\)). This can be used as a simple rule for designing express services: the frequency of an express service \(f_E\) should not fall in the range between \(\hat{f}_E\) and \(P_E\).

This result should come as a surprise since it shows that combinations of express bus services operated under certain frequencies may provide enough transport capacity, but still have several passengers unsatisfied since they will not be able to take the service they prefer. In our example we assumed that bus capacity provides a strict constraint on the number of passengers inside a bus. However, different people may consider different levels of crowdedness as acceptable for their trip. Thus, some passengers may force themselves into the vehicles increasing passenger density beyond what others may consider reasonable. Thus, we would expect this behavior to happen if service frequencies are offered within the infeasibility zone. This may be what happens in very crowded systems as Transmilenio in which express services are widely used and where average densities in some lines often exceed seven passengers per square meter.
4 Determining the optimal solution

The optimal combination of frequencies for the design problem should be the ones that minimize the total social costs, given by:

\[
Z = c_A f_A + c_E f_E + T_A \left( \bar{t}_A + \frac{\theta}{f_A} \right) + T_E \cdot \min \{ C_I; C_{II} \} \quad (7)
\]

In the last expression \( \bar{t}_A \) is the average travel time for users in \( T_A \). The value of \( Z \) will depend on which strategy has the lowest expected cost. This means that \( Z \) can be formulated as:

\[
Z = \begin{cases} 
Z_I = c_A f_A + c_E f_E + T_A \left( \bar{t}_A + \frac{\theta}{f_A} \right) + T_E \left( f_A t_A + f_E t_E + \theta \right), & f_E < f_E^* \\
Z_{II} = c_A f_A + c_E f_E + T_A \left( \bar{t}_A + \frac{\theta}{f_A} \right) + T_E \left( t_E + \frac{\theta}{f_E} \right), & f_E \geq f_E^*
\end{cases} \quad (8)
\]

And the solution of problem \( P \) defined as \( \min Z(f_A, f_E) \) for \( f_A, f_E \in \Omega \) can be obtained by solving independently \( P_I \) \( \min Z_I \) for \( f_A, f_E \in \Omega_I \) and \( P_{II} \) \( \min Z_{II} \) for \( f_A, f_E \in \Omega_{II} \), and then keeping the best of both solutions.
The optimal solution of \( P_I \) is simple to obtain. In fact, the problem is separable for \( f_A \) and \( f_E \), with optimal solution:

\[
\begin{align*}
    f_A^* &= \max\{P_A; \frac{\hat{f}_A}{\hat{f}_E}\}, \\
    f_E^* &= \max\{P_E; \frac{\hat{f}_E}{\hat{f}_A}\}
\end{align*}
\]

Where \( \hat{f}_A \) and \( \hat{f}_E \) are the optimal solutions for the unconstrained version of the problem consistent to the well known square root formula, i.e., \( \hat{f}_A := \sqrt{\theta T_A/c_A} \) and \( \hat{f}_E := \sqrt{\theta T_E/c_E} \).

The first order conditions for \( P_I \) can be stated as:

\[
\frac{\partial Z_I}{\partial f_E} = 0 \rightarrow f_E = \frac{\hat{f}_E + f_A}{f_E} - f_A \tag{9}
\]

A second condition can be derived from subtracting (9) to (10):

\[
\frac{\partial Z_I}{\partial f_A} - \frac{\partial Z_I}{\partial f_E} = 0 \rightarrow f_E = \frac{\tau T_E f_A^2}{\tau T_A f_E - (c_A - c_E) f_A^2} - f_A \tag{10}
\]

The optimal solution for \( P_I \) can be obtained numerically from equations (9) and (10).

5 Conclusions and work in progress

This analysis proves, under some assumptions regarding the rationality of the users, that there can be a range for the frequency of an express service where it leads to overcrowding. This is a relevant and counterintuitive result, since it says that sometimes increasing the frequency of a service can lead to a deterioration of the quality of service for its users. This finding can be interpreted as a public transport version of the well-known Braess’ paradox.

If the frequency of the two services lay within the danger zone, what should we expect from the passengers in \( T_E \) to do? All of them would rather wait for the express service as long as they can board the first arriving bus from this service. However in this zone the demand exceeds its capacity. If we consider a steady state analysis, the demand allocated to a service should not exceed the capacity since otherwise queues would steadily grow. If we assume a more realistic context in which the demand can exceed the capacity for a limited time, we should consider what a user would do if he or she cannot board the first bus. It should be noticed that waiting for the next express bus requires a full average headway so this queuing time takes discrete values (i.e. non-continuous). In any of the two cases, some of the passengers in \( T_E \) should end up taking the regular service when it shows up at the stop. In this paper an equilibrium analysis for this problem will be presented.
Acknowledgements

The authors wish to thank FONDECYT for its funding, project 11,140,443. This study was also supported by the BRT Center funded by the Volvo Research and Educational Foundations (VREF) and the Center for Sustainable Urban Development (CEDEUS), Conicyt / Fonadap / 15110020.

References


