Limited-Stop High-Frequency Service Design: Reducing In-Vehicle Congestion

Andrés Felipe García Albarracín and Daniel Jaramillo-Ramírez

1 Fundación Universitaria Cervantes San Agustín, Calle 77 No 11-63, Bogotá D.C., Colombia
2 Pontificia Universidad Javeriana, Department of Electronics, Cra. 7 No 40B-62, Bogotá D.C., Colombia

Correspondence should be addressed to Andrés Felipe García Albarracín; andrespipe150@gmail.com

Received 20 November 2018; Accepted 7 February 2019; Published 2 July 2019

1 Introduction

A well-known paradox describes a common nonconformity of any transit user: when the user wants to access the system, he/she wants it to have as many stops as possible. When the user is inside the system, he/she wants it to have as few stops as possible before getting off. This paradox is accentuated when cities grow large, dense, or both. A large city will need too many stops on a service making it extremely long and unattractive, sending users to individual and (sometimes) faster transportation options. A dense city will generate too much demand on the service creating in-vehicle congestion and again making transit less attractive. The solution to this problem has been approached mainly from a major related one: the network design, routing, and scheduling problem for an entire transit system. A vast literature record dates from the last 50 years as can be found in Guihaire and Hao [1]. But when cities get denser and larger, in-vehicle congestion and excessively long trips (both in trains and buses) urge for different strategies. Before the modern Bus Rapid Transit (BRT) era, a group of papers including Jordan and Turnquist [2], Ceder and Stern [3], and Furth and Day [4] deal with the problem of heavy demand corridors introducing different bus service strategies. Some of these strategies (actively used today in many cities) are (i) short-turn services, especially useful for corridors with low demand in outer segments; (ii) deadheading, suited for demand imbalance between outbound and inbound directions; and (iii) limited-stop (LS) services (also called express or skip-stop services), mainly adapted for overcrowded routes in central districts or routes with very long All-Stop (AS) services. After the BRT boom (normally credited to Transmilenio in Bogota in 2000) LS services have attracted great attention: they are fundamental in achieving heavy rail-like capacities at a much lower cost. In recent years LS bus services have been successfully implemented in many cities, both for BRT or regular bus services with no right of way. The academic literature though has not been so extensive. Schwarcz [5] and Scorcia [6] propose a framework for the design and performance evaluation of LS services. El-Geneidy and Surprenant-Legault [7] evaluate the implementation of an LS service in Montreal finding that users tend to overestimate the time savings. In the last ten
years, interesting optimization problems were solved to find solutions for the design of LS services. In Sun et al. [8] a scheduling and headway optimization is done for a complete BRT network including LS services. In Leiva et al. [9] an unlimited number of LS routes are allowed to minimize social costs under capacity constraints. Similar conditions are used in Ulusoy et al. [10], but stochastic user behavior is also analyzed. In Chiraphadhanakul and Barnhart [11] only one LS route is used to maximize user welfare (described as in-vehicle time minus waiting time). Capacity constraints are also applied and transfers are not allowed. In Niu [12] a bilevel genetic algorithm (GA) solves a nonlinear program to minimize waiting times and in-vehicle congestion, finding a scheduling and a different set of stops for each departing bus. In Chen et al. [13], headway optimization in LS BRT services is solved with a GA, implemented and evaluated on a BRT corridor in Beijing. Finally, in Zhang et al. [14] a GA minimizes the total trip time adding LS services to those stations where passengers cannot access a congested bus. The work of Larrañ and Muñoz [15] summarizes when LS services are more beneficial: having a high number of trips, long dwell times, and high trip concentration in the OD matrix and when vehicle capacity is more constraining.

Although in-vehicle congestion is one of the most important problems where LS services can be effective, to the best of our knowledge there are no works explicitly evaluating the impact of this strategy or exclusively minimizing in-vehicle congestion. Especially for very high-frequency services (with < 5 minutes headway), sometimes increasing the fleet to reduce congestion is not an option. This might generate bus bunching, excessive on-street congestion, and lack of reliability. It may also be not financially feasible to increase fleet size. Even worse, some transit agencies may compromise the fleet in less used routes, creating a vicious circle for low frequencies and lack of demand.

Let us focus on a very common situation in many cities. A fixed-size fleet operates a very high-frequency service in a traditional AS route. As a result of urbanization processes, the corridor increases its density, and in-vehicle congestion increases together with user dissatisfaction. No more buses can be added (for either budget or traffic restrictions). Can this in-vehicle congestion be alleviated creating an LS service and splitting the existing fleet for both (AS and LS) routes? In this context, given special attention to the passengers’ comfort, the main research questions of this work are as follows:

(i) Is it possible to obtain an important in-vehicle congestion reduction combining LS and AS services with a fixed-size fleet?
(ii) What is the cost to pay in terms of total time?
(iii) Is this gain so important that could lead to a service where all or most of the passengers have a seat?

This set of conditions have an interesting added value on car-free days. Celebrated in more than 1000 cities, their positive impacts can be increased or extended to broader areas of a city if the transit system is able to cope with that day’s peak demand. No expensive infrastructure can be built for only one day a year; hence, properly designed bus routes seem to be a candidate solution.

The rest of this article is organized as follows. Section 2 defines the optimization problem: how to minimize the maximal required capacity for either LS or AS buses. Section 3 describes the model and assumptions used to evaluate LS services in a set of 12000 statistically generated bus corridors. Section 4 presents the solution to the optimization problem finding analytic results on the fleet share for each route and proposing an algorithm to find the optimal set of stops for the LS service. Constraints on the headway are also included. Section 5 presents the results of applying the proposed solution to the set of 12000 corridors in terms of in-vehicle congestion, total travel time, and other important variables. Finally, conclusions are drawn in Section 6.

2. Optimization Problem

As mentioned above, we try to reduce the maximum required capacity (the peak value in the load profile) of a bus-operated public transport system in very low-traffic conditions such as in BRT systems or in a car-free day. We start with some typical corridors where there are only AS buses and then gradually insert LS stations and buses to the system. Unlike other studies where the criteria to introduce these buses and stations are based on the average trip time, we focus on minimizing the required capacity of the buses and hence on improving the passenger comfort. To do so, we have considered the following assumptions:

(i) We look for only one LS service. Throughout the proposed algorithm, the stations where such a service stops and the number of buses it has from the total available fleet is found.
(ii) The total size of the fleet $F$ is fixed and is determined as the number of buses required to achieve a one-minute headway when there are only AS buses.
(iii) Passengers cannot decide which bus they take. We assume they always take the LS service if they have two LS stops between its origin and destination. Otherwise, they always take the AS service. This assumption relies on the fact that as the services are highly frequent, the users might be willing to wait a short time for “their” bus if comfort is guaranteed.
(iv) The buses have infinite capacity. However, as will be shown in Section 5, all values remain in the range of commercial buses.

Additionally, other assumptions from the system’s model are also considered. These are explained in the following section.

3. Model Description

As the objective of the problem consists of reducing the congestion of passengers in the buses, we have focused on modeling the capacity of the buses along the corridor. In what follows, we present the assumptions of the model and the
equations of the capacities of both LS and AS services in terms of the passengers’ spatial distribution along the corridor, the total trip time of a bus in the system, and a fixed demand.

Consider a single-direction bus corridor of length $L$, with a set $\Sigma$ of $S$ stops. Table 1 presents the values of $S$ and $L$ that we have considered for the analysis, providing a good compromise between accessibility and mobility.

In each of the combinations of Table 1, the portion of all passengers that travel from the $i$-th station to the $j$-th station ($j > i$) is modeled by a joint probability distribution of origin-destination, $p(i, j)$. The probabilities of all pairs of stations can be conveniently written in the origin-destination matrix $P$, such that $P_{ij} = p(i, j)$. This way, the sum of entries of the $i$-th row of the matrix corresponds to the portion of all passengers that board at the $i$-th station and, similarly, the sum of the entries of the $j$-th column is the portion of the passengers that alight at the $j$-th station. The joint probability distribution is modeled as a sum of $M$ independent Gaussian random variables or modes (a Gaussian mixture model):

$$p(i, j) = \sum_{k=1}^{M} f_k(i) g_k(j),$$  \hspace{1cm} (1)

where $f_k(i)$ and $g_k(j)$ are obtained from the integral of continuous Gaussian distributions around the $i$-th and $j$-th station, respectively:

$$f_k(i) = \kappa_{f,k} \int_{l_{n}\left(i-3/2\right)}^{l_{n}\left(i+1/2\right)} \frac{1}{\sqrt{2\pi}\sigma_{f,k}^2} \exp\left(-\frac{(x-\mu_{f,k})^2}{2\sigma_{f,k}^2}\right) dx,$$

$$g_k(j) = \kappa_{g,k} \int_{l_{n}\left(j-3/2\right)}^{l_{n}\left(j+1/2\right)} \frac{1}{\sqrt{2\pi}\sigma_{g,k}^2} \exp\left(-\frac{(y-\mu_{g,k})^2}{2\sigma_{g,k}^2}\right) dy,$$  \hspace{1cm} (2)

where $\kappa_{f,k}$ and $\kappa_{g,k}$ are normalization constants and $l_n$ is the distance between consecutive stations. The mean values of the Gaussian distributions are jointly selected from a two-dimensional uniform distribution such that any pair $(\mu_{f,k}, \mu_{g,k})$, with $0 < \mu_{f,k} < \mu_{g,k} < L$, is equally likely. The standard deviations $\sigma_{f,k}$ and $\sigma_{g,k}$ are uniformly distributed between $0.5 \, l_n$ and $1.5 \, l_n$. Figure 1 presents an example of a joint probability function for a corridor with 17 stations and 3 modes, which are at the upper-right triangle of the function domain. The figure also shows the marginal probability functions, which indicate the probability that a passenger departs from or arrives at a particular point along the corridor.

If all of the buses in the system were AS services, a bus passing by the $k$-th station should have a capacity of at least
where $H$ is the headway and $y$ is the number of users per time unit in the corridor. The double sum in (3) corresponds to the portion of all passengers that initiate in the stations before or equal to $k$ (first sum) and terminate at any of the subsequent stations (second sum). This portion multiplied by $y$ gives the rate or number of passengers per time unit that travels by bus at the $k$-th station, whose multiplication by $H$ results in the accumulation of passengers that pass by the $k$-th station between consecutive buses, i.e., the minimum capacity that a bus should have to transport all those passengers.

Now assume that $E$ of the $S$ stations of the system are stops for express or LS services. Those stations are enumerated as $e_k \in \{1, S\}$, such that $\eta = \{e_1, e_2, \ldots, e_E\}$ is the set of all LS stations. With the introduction of LS stations, knowing the next and previous stations where LS buses stop is important to determine the capacity of the AS service at any station. This is due to the fact that LS buses take some of the passengers that AS services used to transport and hence change their required capacity. For a given station $k$, we call the previous station where the express service stops as $a_k$. For a given station $k$, we call the next station where the express service stops again. Observe that $d^k$ can be the $k$-th station itself but $n^k$ is always greater than $k$. We suppose that the express service does not necessarily have to start and finish at the first and last stations of the corridor and, hence, there are some stations where $d^k$ or $n^k$ do not exist. In such cases, we set $d^k = 0$ and $n^k = S + 1$.

Under the assumption that any passenger will use the express service as much as possible during the journey and using the previous definitions, similar expressions to (3) can be written for LS and AS services. In the former,

$$C_{s_k}^E = H^E y \sum_{j=1}^{e_k} \sum_{i=k+1}^{S} P_{ij}, \tag{4}$$

while in the latter

$$C_k^C = H^C y \sum_{i=1}^{k} \sum_{j=k+1}^{S} P_{ij} + \sum_{i=a_k}^{k} \sum_{j=1}^{S} P_{ij}. \tag{5}$$

This last expression converges to (3) when $\eta = \{\emptyset\}$.

The headway between consecutive buses depends on the transit time $t_{in}$ from one station to the next one and the service time at each station $t_s$. Here we assume that the stations are equally spaced and hence $t_{in}$ is constant for any given $S$ and $L$. With such conditions, the LS and AS headway in (4) and (5) are

$$H^E = \frac{(e_E - e_1 + 1) t_s + 2(e_E - e_1 + 1) t_{in}}{F_E}, \tag{6}$$

$$H^C = \frac{S t_s + 2S t_{in}}{F_C}, \tag{7}$$

respectively, where $F_E$ and $F_C$ are the number of buses in the LS and AS service, such that the sum $F_E + F_C = F$ is the constant number of identical vehicles in the corridor. The total transit time appears twice in each of (6) and (7) to include the return trip, but $t_s$ appears only once, as we assume deadheading of all buses in the opposite direction. Final expressions of the capacities for LS and AS services are obtained by replacing (6) and (7) in (4) and (5). These two expressions are used in the solution of Section 4.

Regarding the values that some variables take in our model, we have assumed the following:

(i) The number of users per time unit in the corridor, $y$, is dependent on the corridor length as we have supposed that one passenger arrives per hour, per direction, per meter. Thus, the demand will be $y = 5000 \text{ pax}/\text{h/dir}$ in a $5 \text{ km}$ corridor.

(ii) The service time $t_s$ is constant for every station and equals to $20 \text{ s}$, which is valid under the assumption of high-frequency service.

(iii) We obtain the transit time $t_{in}$ from some standard calculations of a typical BRT system.

### 4. Proposed Algorithm

The algorithm we have proposed selects the stations and number of buses of the express services by trying to take the largest amount of passengers from the AS line and along the longest distance. To do so, we perform Hadamard multiplication (entry by entry) between the probability matrix $P$ with a matrix $Q$ containing distance factors between stations. Because the stations are equally spaced, the matrix $Q$ basically consists of a linear series along its antidiagonals as shown in

$$Q = \begin{bmatrix}
0 & 1 & 2 & \cdots & S-1 \\
1 & 0 & 1 & \cdots & S-2 \\
2 & 1 & 0 & 1 & \cdots & S-3 \\
\vdots & \vdots & \vdots & \ddots & \ddots \\
S-1 & \cdots & \cdots & \cdots & 0
\end{bmatrix} \tag{8}
$$

The entries of the product $PQ$ are then sorted in descending order to obtain the ordered set $\Theta$ of origin-destination stations which are suitable to be selected as stops of the LS buses. Because each entry of the product $PQ$ gives two stations, the origin is added first to the set $\Theta$ and then the destination. This way, $\Theta_1$ and $\Theta_2$ contain the origin and destination of the first pair, $\Theta_3$ and $\Theta_4$ the origin and destination of the second one, and so on. The set $\Theta$ is the basis for the subsequent iterations of the algorithm.

Each iteration $l$ of the algorithm starts by adding a new station from the set $\Theta$ to the set of express stations $\eta$, as shown in Figure 2. Notice that $\eta^l$ means the set $\eta$ at the $l$-th iteration and $\Theta_k$ is the $k$-th element of the $\Theta$ set. This occurs in every iteration, except in the first one, where $\eta$ contains the first two stations of $\Theta$.

With the set $\eta^l$, the next step consists of finding the optimum values of $F_E$ and $F_C$ that minimize the maximum of
the capacities among all the stations of the system. Observe from (4)-(7) that the capacities can be written as
\[
C_k^E = \frac{\alpha_k}{F_E}, \quad (9)
\]
\[
C_k^C = \frac{\beta_k}{F - F_E}, \quad (10)
\]
where \(\alpha_k\) and \(\beta_k\) are station-specific factors that contain the transit time of LS and AS services, the passenger’s rate \(\gamma\) and the sum of some entries of the probability matrix \(P\). The denominator in each case is the number of buses available for each service, which is constant for every LS or AS capacity. Notice that the maximum of the capacities of LS and AS services occurs at the maximums of \(\alpha_k\) and \(\beta_k\), regardless of the values of \(F_E\) and \(F_C\). Let us call the indexes of those maximums \(k^*\) and \(k^\dagger\), respectively. Therefore, the problem of minimizing the maximums of the capacities among all the stations and services is reduced to
\[
F_E^* = \arg \left\{ \min_{F_E} \left[ \frac{\alpha_{k^*}}{F_E} + \frac{\beta_{k^\dagger}}{F - F_E} \right] \right\}. \quad (11)
\]
The first expression in the minimization of (11) strictly decreases with \(F_E\), while the second term strictly grows with \(F_E\). Hence, the minimum occurs at the intersection of both terms, i.e., when the maximum of the capacities of LS and AS services is the same, or as close as possible considering that \(F_E\) and \(F_C\) can only take discrete values. This allows us to obtain closed forms for the optimum buses in each service:
\[
F_E^* = \text{round} \left( \frac{\alpha_{k^*}}{\alpha_{k^*} + \beta_{k^\dagger}} \right), \quad (12)
\]
\[
F_C^* = F - F_E^*, \quad (13)
\]
which are saved in each iteration along with the optimum value of the capacity, \(C^*(\eta_l)\).

Once the optimum values are obtained, the solution is considered feasible as long as any of the following conditions do not occur:

(i) At least one of the headways exceeds a maximum value \(H_{\text{max}} = 5\) min, making unfeasible our assumption that every passenger would take an express service as much as possible (passengers would not find attractive waiting long time intervals when transferring from one bus to the other).

(ii) At least one of the headways is lower than a minimum value \(H_{\text{min}} = 0.5\) min, which cannot be implemented in practice.

(iii) Either \(F_E\) or \(F_C\) is zero.

If one of the three conditions is satisfied, the algorithm discards the current unfeasible results. After such a decision, the algorithm proceeds to the next iteration up to \(I = S\), which is equivalent to \(I = 0\). Observe that, including the case with only AS stations, the algorithm evaluates up to 5 possible scenarios or distributions of LS stations.

Finally, the minimum value of the capacity in all the evaluated scenarios is found, which also gives the gain of the algorithm when comparing to the case with only AS services.

5. Results

5.1. Efficiency and Overall Gains. Overall, the algorithm’s efficiency is greater than 90% in all of the scenarios studied in this paper. We have assumed that the algorithm is efficient whenever the maximum of the initially required capacity is reduced. The 10% of inefficiency occurs when it does reduce the capacity but violates any of the constraints of Section 4 in every iteration, or because it simply fails at reducing the capacity. The latter, however, only occurs in average at 0.55% of the simulated cases.

As shown in Figure 3, the algorithm is especially efficient when the number of modes is small. Observe that when \(M = 1\), the efficiency is in average 99.55%, while it is 96.17% when there are two modes in the corridor and reduces to 91.63% when \(M = 3\). The same occurs with the mean capacity gain, which we define for each iteration as
\[
\delta_C = \frac{C(1) - \min_l [C(l)]}{C(1)}, \quad (14)
\]
where \(C(1)\) is the largest capacity when there are only AS buses and \(C(l)\) the largest capacity at the \(l\)-th iteration. The
Table 2: Average efficiency and capacity gain with different modes along the corridor.

<table>
<thead>
<tr>
<th>Number of modes $M$</th>
<th>Average efficiency</th>
<th>Average capacity gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99.55%</td>
<td>35.06%</td>
</tr>
<tr>
<td>2</td>
<td>96.17%</td>
<td>20.11%</td>
</tr>
<tr>
<td>3</td>
<td>91.63%</td>
<td>15.04%</td>
</tr>
</tbody>
</table>

The mean value is obtained by averaging over all the simulations of a specific corridor length and number of modes. When plotted against the number of modes, the mean gain of the capacity decreases with the number of modes. When there is a single mode along the corridor, the gain can be up to 39.88%, while it is at most 16.39% when it is a double-mode corridor. The dependencies of the mean gain and the efficiency with the number of modes are summarized in Table 2.

The reduction in the capacity gain with the number of the modes might be a result of the fact that whenever there are two or more modes, the LS service does not carry passengers between a pair of stations exclusively. With several modes, the distinguishing characteristic of the express service is somehow lost as it is almost impossible to solve the peak demands of all modes simultaneously. Each time a mode is introduced, the performance of the other solutions is affected.

Regarding the mean trip time of the corridors, Figure 3 shows that it is always affected by the algorithm. When an express service is included in the system, it can reduce the trip time of its users and, simultaneously, it increases the trip time of uncovered users as the number of AS buses is reduced. The latter effect is always more significant than the former, which explains the right chart of Figure 3. Additionally, notice the existence of a strong dependency between the time losses and the corridor length such that, in general, shorter corridors present larger time losses. This is because including express buses in the system only eliminates the service times of the stations where they do not stop but does not remove the transit time between stations. In large corridors, the transit time is more significant than the service time and, hence, minor changes in the latter do not affect a lot the total trip time.

The capacity gain also varies with the corridor length. The dependency, however, is not as strong in the mean values as in the case of the time lengths. Here, it is the entire distribution which changes as illustrated in Figure 4. In general, the distribution of the capacity gain in a large corridor has a longer tail and a longer mean than in a short corridor.
5.2. Minimum Required Capacity. We have also plotted a scatter diagram between the capacity gain and the resultant minimum required capacity. This chart is depicted in Figure 5, where we have included some results of all the simulated scenarios. In agreement with the previous analysis, corridors with fewer modes present also larger capacity gains. The gain sometimes can be up to 60% and in 26.67% of the simulated cases, its reduction is greater than 33%. Those cases shaded in gray in Figure 5 are especially important as usually around one-third of the capacity of any bus corresponds to seated passengers. A reduction of the required capacity by 33% implies that those buses, which traveled completely full in the only AS services scenario, now have as many passengers standing as sitting. Therefore, most of the passengers can now take a seat along those segments of the corridor whose required capacities are smaller than the largest one.

The results of Figure 5 are bounded by some inclined lines which represent the maximum capacity of a corridor. Recall from Section 2 that the initial headway of AS buses was fixed to 1 min and that we assumed that one passenger per meter per hour arrives in the system. With those conditions, the minimum required capacities go up to 83, 167, 250, and 334 when the corridor lengths are 5, 10, 15, and 20 km. The inclined lines of Figure 5 correspond to the variation of those boundaries with the capacity gain. Interestingly, some services that initially exceeded the capacity of biarticulated buses can now be comfortably served with articulated buses.

5.3. Parameters of the LS Services. The required capacity of a corridor is mainly reduced when the headway of LS services can be reduced as much as possible. This occurs when most of the passengers concentrate around a few stations and introducing an express service for them takes out a large portion of the passengers of AS services. The results are summarized in Figure 6 for the scenario when the corridor is 15 km long and has two modes. Additionally, observe that the largest gain is obtained when the number of stations coincides with twice the number of modes of the corridor. That is, the largest performance is obtained when it is enough to stop express buses in those stations with the largest product between distance and joint probability factor.

6. Conclusions

We have proposed a method to reduce the required capacity of a bus-operated transport system by introducing express buses. The criteria to determine the parameters of the express service are based on selecting the largest entries of the joint-probability matrix Hadamard-multiplied by a factor matrix
Figure 5: Scatter diagram of the minimum required capacity and the capacity gain.

Figure 6: Scatter diagram of the capacity gain against the optimum number of LS stations and the LS headway. Blue dots correspond to capacities above the mean, while red points are below it.
accounting for the distance between stations. We tested the algorithm on random corridors of 5, 10, 15, and 20 km with 1, 2, or 3 modes; each combination simulated 1000 times. Our results can be summarized as follows:

(i) The algorithm does reduce the required capacity of the buses along the corridor in more than 90% of the cases. Most of the cases where it does not reduce the capacity, the failure results from the constraints of the maximum and minimum headways that we imposed, typical of a high-frequency service. Only in less than 0.6% of the cases, the algorithm does not reduce the capacity at all.

(ii) The largest efficiency and gain of the algorithm are obtained when the corridor has a single mode. Overall, the larger the number of modes, the less the efficiency and the capacity gain.

(iii) The side effects on the mean trip time are more notorious on short corridors, where changing some few stops is more representative in the total trip time.

(iv) The algorithm achieves gains higher than 33% in more than 26% of the cases, which is especially important to reduce passenger’s congestion in the buses.

(v) In general, the performance of the algorithm is better when the conditions of the corridor (i.e., the joint probability matrix) allow a solution with a small number of stations and a short LS headway.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References
