CASPT 2015

When to provide express services for buses?

Homero Larrain · Juan Carlos Muñoz

Abstract Express bus services have been implemented in many public transit systems around the world to the benefit of users as well as operators. However, few studies have been published on methodologies for designing such services and there is little clarity regarding what factors in a bus corridor make it a good candidate for a successful application of this transport option. The present article reports on an experiment in which an express service design algorithm developed for the purpose optimized almost 1,000 different scenarios and a model formulated for use with the algorithm estimated the benefits of operating express services as a function of a series of corridor attributes. An analysis of the calibrated model demonstrated that a corridor's potential for beneficial express services increases with increasing dwell times, number of trips, concentration of trips in few origin-destination pairs and critical arc load, and decreasing travel and wait time values and vehicle capacity. An analytic expression was also derived that estimates the percentage savings obtainable from the implementation of an express service as opposed to a traditional all-stop service.

Keywords: Express Services · Bus Rapid Transit · Public Transport · Network Design

1 Introduction

Express bus services have existed in many cities around the world for more than 50 years and are highly popular among operators as well as users. For operators such

---

Homero Larrain
School of Transportation & Logistics Engineering, Pontificia Universidad Católica de Chile,
Santiago, Chile
Email: homero@ing.puc.cl

Juan Carlos Muñoz
School of Transportation & Logistics Engineering, Pontificia Universidad Católica de Chile,
Santiago, Chile
Email: jcm@ing.puc.cl

services offer a more efficient utilization of bus fleets and therefore lower costs per passenger, while for users they provide rapid transport to the desired destination. Yet there is still no consensus in the literature on how best to design express lines or what conditions they generate the most benefits under.

Since the turn of the 21st century, express bus operations have multiplied at a rate never before witnessed. The phenomenon is part of a broader international trend towards strengthening public transport and discouraging the use of private vehicles, compelling bus operators to constantly innovate and improve their services. This tendency has given particular impetus to the adoption of systems known as bus rapid transit or BRT. The New York Metropolitan Transportation Authority (MTA) has defined BRT as "a flexible, integrated, high performance transit system with a quality image and a strong identity", combining the speed, reliability and other virtues of rail-based systems with the flexibility of buses (www.mta.info). In the words of Lloyd Wright (2003), BRT is "an alternative between poor service and high municipal debt. … [It] can provide high-quality, metro-like transit service at a fraction of the cost of other options". According to Global BRT Data (www.brtdata.org), there are an estimated 168 cities worldwide that have implemented some sort of BRT along its streets. Some 80% of these systems have been inaugurated since the year 2000 and almost half of them have been in operation for no more than 7 years.

There are several elements commonly associated with BRT systems. These typically include segregated bus lanes, centralized fare collection, modern bus stops with off-board fare payment and expedited boarding, a uniform vehicle image, and traffic signal priority (Levinson, 2002). To achieve capacity levels comparable to rail systems, however, a key measure is the provision of express services. These can be operated with BRT systems that have passing lanes at stops. This feature allows Bogota's Transmilenio, for example, to transport 48,000 passengers per hour in each direction, thereby attaining a commercial speed of almost 30 km/hr (Hidalgo and Munoz, 2014).

The definition of an express service, also known as a limited-stop service, can be stated in general terms as a service that visits a subset of stops along a given corridor. As suggested above, these services constitute an operating strategy that can be beneficial to all stakeholders. For passengers, the relatively small number of stops implies a reduction in travel times, due either to the direct time savings from not stopping as frequently or the possibility of taking alternative route segments that are shorter or faster. For system operators, the reduction in route cycle times means more efficient vehicle utilization and therefore lower operating costs. Furthermore, the higher speeds allow operators to provide a given level of transport capacity with a smaller fleet, thus giving them the flexibility to choose between lowering costs while maintaining capacity or improving frequency and comfort. But if express services are to be a genuinely attractive alternative, these advantages must be enough to offset the increases in wait times and transfers for users at skipped stops. In other words, these services will not always be a net benefit.

The central question the present study attempts to answer is what attributes of a bus corridor make it attractive for the implementation of an express service. The ultimate objective is to develop an indicator for the attributes that estimates the general benefits of implementing this type of service along a given corridor. Our methodology for answering this question will consist in constructing a series of scenarios with different characteristics, designing a solution using express services for each scenario and solving an optimization problem for each one, measuring the benefits generated by these solutions, and finally, calibrating an econometric model to estimate the benefits obtainable as a function of various corridor attributes. The model will thus be an estimator of the potential of express services for different corridors, from which some simple rules can be derived for identifying opportunities to implement them.

1.1 Literature Review

According to Desaulniers and Hickman (2007), the first study to address the express service design problem was Jordan and Turnquist (1979), which optimizes the operation of a corridor on which all trips are attracted to (generated at) the first (last) stop. This model was generalized by Furth (1986), who studied the problem for the case of bidirectional and branching corridors while maintaining the same demand structure.

Some authors have evaluated the performance of express services through case studies. One such effort is the work of Ercolano (1984), who surveyed users of express services in the New York borough of Manhattan. Another example is Silverman (1998), who conducted an analysis using data from 25 New York City services. Both investigations found the measure to be effective and reported positive benefits for operators as well as passengers. Tétreault and El-Geneidy (2010) proposed an express service for Montreal with four different scenarios for selecting stops, prioritizing those where demand is greatest. El-Geneidy and Surprenant-Legault (2010) evaluated the implementation of this proposal and detected both travel time savings and a positive perception of the service on the part of users.

A number of recent works have developed methodologies for designing corridor express services and have performed various sensitivity analyses of the results. Sun et al. (2008) formulates an integer programming model that optimizes the configuration and frequency of three services (regular, express and zonal) assuming passengers minimize their itinerary by choosing a single service without considering the possibilities offered by common (i.e., overlapping) lines. The authors identified trip volume and low operating speed as indicators that express service should be provided. Chen et al. (2012) devise a model that minimizes social costs for a corridor served by regular and express services and present a case study for the Beijing BRT system, but include no sensitivity analysis.

Chiraphadhanakul and Barnhart (2013) propose an incremental algorithm for designing an express service operating alongside a regular service. Their methodology assumes that some users prefer to wait for express service, the proportion so doing depending on the time savings it affords, but takes no account of service frequency. After analyzing 178 bus routes the authors conclude that express services should be implemented when average trip length and the ratio of the value of travel time to that of wait time are relatively high. They also contradict the observation of Sun et al. (2008) (and later Larrain, 2010a), finding that demand volumes are correlated with total user welfare only in an absolute sense.

Leiva et al. (2010) develop a model for designing express services on a limited-capacity public transit corridor. Their proposed formulation minimizes social costs (i.e., operator plus user costs) for a predefined set of candidate express services. The model's results are consistent with rational behaviour on the part of users (who minimize their itineraries considering the existence of common routes). This is so even in the presence of vehicle capacity constraints, which are handled by increasing the lower bounds on service frequency until the offer is consistent with the desired behaviour. Larrain et al. (2015) address the problem of generating zonal services (a special case of express services) to feed the Leiva model and proposes various other modifications to it including an improved capacity adjustment heuristic.

The Leiva model is applied in Larrain et al. (2010a) to analyze the potential of express services in different corridor scenarios. In the case of a corridor in Santiago, Chile, the model finds that express services would generate total social cost savings on the order of 10%. In this as well as other cases reported by the authors, much of the savings are derived from a small number of different services. The indicator found to be best correlated with the savings obtained is the average trip length, although it was also observed that express services generate greater savings percentages at higher demand levels.

Larrain (2010b) extends Larrain et al. (2010a) to include an analysis of the effect of specific origin-destination pairs with high trip concentrations. Using the trip matrix's variability coefficient as an estimator of this attribute (the higher the coefficient, the more homogeneous the trip dispersion, that is, the lower the trip concentration), the authors conclude that the benefits of the express services increase when the trips are more concentrated in a small number of O-D pairs.

1.2 Conceptualization of the question

To develop an answer to the central question posed by this study as set out above, we must first establish (1) the set of possible variables that influence the impact express services have on the efficiency of a BRT system, and (2) how the performance of an express service solution should be measured. In what follows, therefore, we specify the explanatory variables that were included in the experiment carried out for this study and the performance indicator used for evaluating the candidate solutions.

The explanatory variables were chosen partly on the basis of the results reported in the works discussed above in the literature survey, and partly based on our own experience working with express service design algorithms. These variables can be classified into three groups: demand characteristics, operating characteristics and relative weights of the cost components. The three groups are taken up in turn below.

1. Demand characteristics: The experiment we conducted used trip matrices for a variety of trip volumes, average trip lengths and trip concentration levels. The passenger load profile shape as such was not included as an explanatory variable given that in Larrain et al. (2010a) its effect was shown to be unclear and attributable in large part to average trip length. Also left out was demand imbalance as this factor refers to highly specific cases already studied in the literature (deadheading) and can in any case be considered as a form of trip concentration. A new factor in the analysis is the critical arc load, incorporated through the minimum all-stop service frequency. By comparing this factor to the unrestricted optimal frequency we can determine the degree of vehicle overcrowding.
2. Operating characteristics: One of the most important factors in our experiment is bus stop dwell time. The potential savings in travel time costs depend directly on this phenomenon. Another variable included in the present study is vehicle capacity, which has a direct impact on vehicle overcrowding. Earlier analyses have already shown this factor to be a good predictor of express service potential. A third variable we incorporated is corridor length in terms of the number of bus stops. Other operating and topological parameters such as speed, distance between stops and therefore travel time between stops were excluded from our experiment, however. From the passenger's point of view, travel time between stops is a constant that cannot be avoided and therefore has no effect on service optimization, while from the operator's standpoint, the effect of varying distances (or times) between stops is similar to that of modifying operating cost parameters and is accounted for in that form in our modelling, as will be explained in the next item.
3. Relative weights of the cost components: The social costs to be optimized by our proposed model have four components: operator costs, in-vehicle travel time costs, wait time costs and transfer costs. Each of these components has one or two associated parameters that determine its weight in the model's objective function. Operator costs depend on the unit distance operating cost and the unit time operating cost. In-vehicle travel time is weighted by the travel time value while wait time is weighted by the wait time value. Since we will only consider medium and high frequency services, we assume that wait times are inversely proportional to the service frequency and directly proportional to the parameter λ. This parameter is just the average proportion of the average interval between buses that a passenger must wait and represents the random factor in bus arrivals. However, since it always appears in our formulation multiplied by the wait time value (as will be see below), there is little point in subjecting it to a separate sensitivity analysis. Finally, transfers are weighted by a transfer cost parameter.

The performance (quality) indicator used to evaluate the candidate express service solutions in a given scenario can be specified in various ways. In this study we will use an indicator we call corrected percentage savings, defined as the percentage difference between the corrected costs of the solutions for a given scenario with and without express services. The corrected cost of a solution is defined as its social cost less the fixed element of travel time users must incur. The detailed specification of these costs will be given in Section 3.

The remainder of this article is organized into four sections. Section 2 introduces the algorithm used to design the express services to be evaluated in our experiment; Section 3 reports on the experiment, detailing the construction of the scenarios to be optimized, the calculation of the performance indicator and the calibration of the multiple linear regression model for estimating the indicator values; Section 4 contains various analyses using the calibrated model for answering the central question posed in this study; and finally, Section 5 presents our conclusions.

2 Express service design algorithm

To evaluate the effects of the different variables on express service performance we develop a tool that optimizes the design of express services along a corridor for the different scenarios to be evaluated. The tool consists of an with two modules that feed back to each other, one generating candidate services and the other optimizing service frequencies.

The algorithm assumes that the corridor to be optimized has $2n$ stops (n in each direction) and that stops 1 to n in the outbound direction coincide geographically with the inbound stops $n+1$ to $2n$. It is also assumed that demand between stops is fixed and known and that, to facilitate optimization, users do not make transfers and the buses are all of the same vehicle type.

2.1 Frequency optimization

The frequency optimization problem consists in determining, for a given corridor and set of candidate services to be operated on it, what is the vector of frequencies that minimizes social costs. The vector may have various zero elements, indicating that some available services may not be used. In this context a service is defined as a sequence of nodes to be visited in a given corridor direction. Thus, the need to work with a single vehicle type.

The notation used in the rest of this article is summarized in Table 1.

Table 1: Notation.

Notation	**Description**
$\boldsymbol{\mathcal{L}}$	Set of candidate services to be evaluated.
$\boldsymbol{\mathcal{W}}$	Set of corridor origin-destination (O-D) pairs.
$\boldsymbol{\mathcal{N}}$	Set of corridor nodes, $\lvert\mathcal{N}\rvert = 2n$.
$\boldsymbol{f_l}$	Frequency of service $l \in \mathcal{L}$.

$\boldsymbol{f}$	Vector of service frequencies, $\{f_l\}$.
$\boldsymbol{g_l^w}$	Frequency of service $l \in \mathcal{L}$ perceived as attractive for O-D pair w; equal to 0 or f_l.
$\boldsymbol{g}$	Vector of attractive service frequencies $\{g_l^w\}$.
$\boldsymbol{c_l}$	Cost associated with service $l \in \mathcal{L}$ per frequency unit (note that c_{AS} is for all-stop service).
$\boldsymbol{c_L}, \boldsymbol{c_T}$	Variable operating costs per unit of distance and time, respectively.
$\boldsymbol{T_w}$	Demand for trips between O-D pair $w \in \mathcal{W}$.
$\boldsymbol{T}$	Total corridor demand, i.e., $T = \sum_{w \in \mathcal{W}} T_w$.
$\boldsymbol{P}$	Critical arc load of a corridor.
$\boldsymbol{\lambda}$	Wait time parameter.
$\boldsymbol{\beta}$	Demand volume amplification parameter.
$\boldsymbol{t_l^w}$	Travel time for service $l \in \mathcal{L}$ between O-D pair $w \in \mathcal{W}$.
$\boldsymbol{\theta_{WT}}$	Wait time value.
$\boldsymbol{\theta_{TT}}$	In-vehicle travel time value.
$\boldsymbol{b}$	Vehicle capacity (in passengers).
$\boldsymbol{O(w)}, \boldsymbol{D(w)}$	Origin and destination nodes, respectively, for O-D pair $w \in \mathcal{W}$.
$\boldsymbol{O(l)}, \boldsymbol{D(l)}$	Initial and terminal nodes, respectively, of a service l ∈ L.

The service frequencies along a corridor with no transfers are optimized by the following model:

$$Min\left[TC(\boldsymbol{f},\boldsymbol{g}) = \sum_{l \in \mathcal{L}} c_l f_l + \theta_{WT} \sum_{w \in \mathcal{W}} T_w \frac{\lambda}{\sum_{l \in \mathcal{L}} g_l^w} + \theta_{TT} \sum_{w \in \mathcal{W}} T_w \frac{\sum_{l \in \mathcal{L}} t_l^w g_l^w}{\sum_{l \in \mathcal{L}} g_l^w}\right] \tag{1}$$

Subject to

$$g_l^w \leq f_l, \qquad \forall l \in \mathcal{L}, \forall w \in \mathcal{W} \tag{2}$$

$$g_l^w \geq 0, \qquad \forall l \in \mathcal{L}, \forall w \in \mathcal{W} \tag{3}$$

$$\sum_{l:O(l)=i} f_l + \sum_{l:O(l)=2n+1-i} f_l = \sum_{l:D(l)=i} f_l + \sum_{l:D(l)=2n+1-i} f_l, \qquad \forall i = \{1, \dots, n\} \tag{4}$$

This formulation is a variation on the one presented in Larrain et al. (2015), which was modified to allow for short turn services (i.e., services that cover only a part of a complete corridor). The model minimizes the social costs of the corridor as a function of the service frequencies $\boldsymbol{f}$ and $\boldsymbol{g}$. The social cost function (1) that is optimized is the sum of the following three terms:

1. Operator costs (first term). Assumed to be proportional to service frequency. Operating costs per frequency unit of service l, denoted c_l, are estimated as follows:

   $$c_l = c_L cl_l + c_T ct_l \tag{5}$$

   where parameters cl_l and ct_l represent the cycle distance and cycle time of service l. This definition implies that the operation costs per frequency unit for an express service are lower than those for a normal service because the former has a shorter cycle time.
2. Wait time costs (second term). Obtained assuming that the wait time for an O-D pair is the average headway of attractive services multiplied by the constant λ described earlier.
3. Travel time costs (third term). For each O-D pair, based on the average travel time weighted by the corresponding frequency.

Constraints (2) and (3) define the relationship between frequencies and attractive frequencies. Constraint (4) is a flow conservation condition for buses at each stop, ensuring that for each node i and its corresponding node in the opposite direction $2n + i - 1$, the sum of the frequencies of services terminating their run must be equal to the sum of those for services initiating their run.

Note that model (1)–(4) does not take into account vehicle capacity. Imposing a direct capacity constraint may result in solutions whose user assignment does not satisfy the user equilibrium conditions. To address this problem, Leiva et al. (2010) propose an algorithm, improved in Larrain et al. (2015), in which the frequency of the service with the greatest capacity deficit is increased iteratively until a solution is reached that does not violate vehicle capacity. A variation on this algorithm in which express services are prioritized at the time frequencies are increased is proposed in Larrain (2013). In the present study these two heuristics will be referred to as CAP and $CAPe$, respectively.

2.2 Generation of services

The service generation problem consists, as its name suggests, in generating services that would help construct better solutions or improve existing ones for the frequency optimization problem. In this context a service is defined as a sequence of stops to be visited in a given corridor direction.

A series of heuristics for generating potentially beneficial express services along a corridor is presented in Larrain (2013). The main characteristics of these heuristics are set out in Table 2.

Table 1: Heuristics for generating express services along a corridor.

Name	Capacity constraints	Short or zonal	Existing services	Description
GEN_1	No	No	No	Generation of initial services: Using a greedy heuristic, chooses an all-stop service and sequentially eliminate some of them on a local optimality criterion, thus generating an initial list of express services.
GEN_2	No	Yes	No	Generation of short services: Builds a list of short turn services.
GEN_3	No	Yes	No	Generation of zonal services: Builds a list of zonal services that visit a complete set of stops in the initial and final segments or zones of a route while skipping a large number of consecutive stops constituting the middle zone.
GEN_4	No	No	Yes	Generation of additional services: Builds a series of new express services for a corridor from a given solution. The services are created by adding stops to an initial base service.
$GENc_1$	Yes	Yes	No	Short service considering capacity: Builds a solution that satisfies the capacity constraint using a short service and an all-top service in the critical direction of the corridor.
$GENc_2$	Yes	Yes	No	Zonal service considering capacity: Builds a solution that satisfies the capacity constraint using a zonal service and a regular service in the critical direction of the corridor.
$GENc_3$	Yes	Yes	No	Zonal service considering capacity: Builds a solution that satisfies the capacity constraint using a zonal service, a short service and an all-top service in the critical direction of the corridor.

Note that Table 2 classifies the services by three different criteria. The second column indicates whether the various heuristics consider vehicle restrictions in the design. For those that do not, the list of services they generate can be added to the set of services L of the frequency optimization problem, leaving capacity to be corrected later using CAP or $CAPe$. The heuristics that do consider capacity directly deliver a solution that uses a reduced number of services but satisfies the capacity constraints.

The third column of Table 2 indicates whether or not the services generated by the heuristics are in either of the two specific formats: short (all-stop service) or zonal (skipping a set of consecutive stops in the corridor's middle segment). The zonal service heuristics with and without capacity constraints are described in detail in Larrain et al. (2015).

Finally, the fourth column of the table indicates whether the heuristics are based on a pre-existing solution for the corridor.

2.2 The algorithm

The service generation heuristics just described are combined with the frequency optimization model into a two-module algorithm for generating express services that was first proposed in Larrain (2013). This algorithm operates in three stages. The first stage begins with an initial solution for an all-stop service. In the second stage, heuristics are applied to generate services for the case where there is no overcrowding, that is, without considering vehicle capacity. If this solution does not satisfy the vehicle capacity constraints, the algorithm proceeds to a third stage in which the capacity restrictions are satisfied. The three stages are described more fully below.

1. Initial solution with all-stop service.
    a. Optimize the operation of an all-stop service modelled for both directions. The optimal frequency (which is necessarily the same in both directions) is given for this case analytically by the following formula:

$$f_{AS}^{*} = \max\left(\frac{P}{b}, \sqrt{\frac{\theta_{WT}\lambda T}{c_{AS}}}\right) \tag{6}$$

    This expression chooses the maximum value of two terms. The first term is the frequency that ensures vehicle capacity is sufficient to carry the critical arc load while the second term represents the optimal frequency if capacity is disregarded (Mohring, 1972, Jansson, 1980). The c_{AS} term is the all-stop service operating cost and is given by (5) in Section 2.

2. Solution to problem without considering capacity.
    a. A priori service generation
        i. Apply GEN_2 to identify the set of short services $\mathcal{L}_C^*$ that could be attractive. Also include in L_C^* all-stop services from terminus to terminus for both directions. Do $\mathrm{L} \leftarrow \mathrm{L}_\mathrm{C}^*$.
        ii. Apply GEN_1 using as a base each element of $\mathcal{L}_C^*$ that has three or more stops. Incorporate the new services thus obtained into set L.
        iii. Apply GEN_3, incorporating the newly generated zonal services into L.
        iv. Optimize the frequencies using model (1)–(4) (without capacity adjustment) and eliminate the services with zero frequencies from L.
    b. Generation of additional services.
        i. $\forall (i,j) \in \mathcal{N} : \exists l \in \mathcal{L} : O(l) = i \wedge D(l) = j,$ apply GEN_4 using as a base the service that connects the nodes i and j with no intermediate stops.

    ii. Optimize frequencies using model (1)–(4) (without capacity adjustment) and eliminate the services with zero frequencies from L.

If the solution at this point does not satisfy the vehicle capacity constraints, proceed to the third stage:

3. Solution to problem considering capacity.
    a. Store the current solution to the problem without capacity constraints.
    b. Starting from this solution, apply algorithms CAP and CAP_e storing both solutions thus obtained.
    c. Solve the problem using $GENc_1$, $GENc_2$ and $GENc_3$, storing the solutions thus obtained.
    d. Compare the solutions obtained in 3.b, 3.c and 1.a, and choose the best one.

3 The experiment

In this section we report on the experiment carried out as part of this study to establish which attributes of a bus route corridor determine the benefits to be gained from express services. To this end we constructed a series of scenarios incorporating different values for the explanatory variables proposed in Section 1. The express service design algorithm presented in Section 2 was then applied to each of these scenarios. The results so generated formed the basis for the calculation of the (corrected) percentage savings on costs afforded by each scenario as measured in relation to an optimized all-stop service. Three linear regression models were sequentially calibrated to determine the effects of the explanatory variables on the performance indicator. The following paragraphs describe the construction of the scenarios to be evaluated, the derivation of the corrected percentage savings and the calibrated models; the results will be analyzed in Section 4.

3.1 Definition of the scenarios

Due to the many variables involved, a number of simplifications were made in order to limit the quantity of scenarios in the experiment and facilitate their solution. One such measure, already mentioned in Section 3, was to assume the cost of transfers is sufficiently high that system users consider only direct routes. Also, as noted in Section 1 the demand imbalance and load profile shape factors were omitted.

Yet another simplification was to vary the two parameters defining operator costs (c_L and c_T) simultaneously, implying that the ratio between them remains constant. This meant that the experiment was able to measure the effect of operator costs as a whole but not—at least directly—broken down by cost component. Finally, since operator costs and the travel and wait time values are the relative weights of the objective function terms, we opted to fix one of their values, namely, the travel time θ_{TT}.

To generate the experimental scenarios, we defined four base trip demand matrices and a series of parameters to be varied, the latter consisting of wait time value, operator cost, total demand, dwell time and vehicle capacity. The matrices were generated from passenger load profile data on two bus corridors in the city of Santiago, Chile: Avenida Pajaritos and Avenida Grecia. For both corridors the load was distributed among 10 and 20 stops per direction, thus obtaining the four matrices shown in Figure 1.

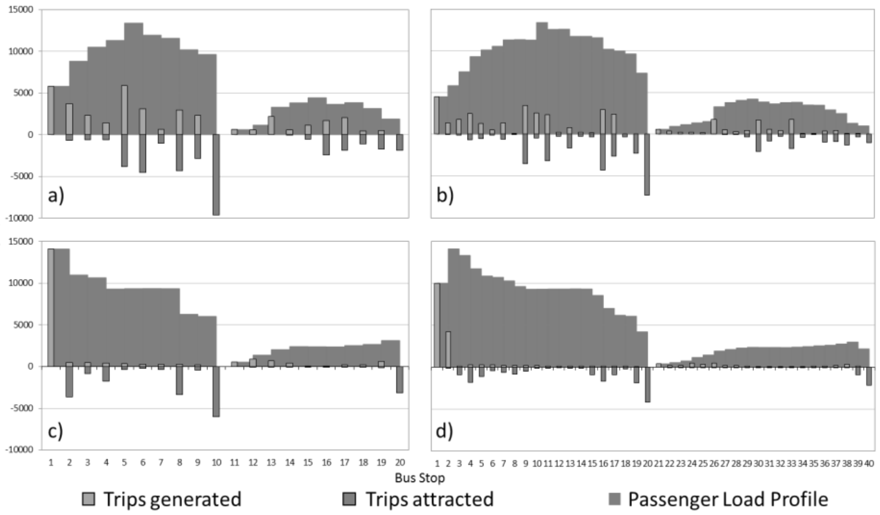

Figure 1: Load profiles of base trip matrices.

The figure shows the trips generated and attracted as well as the load profiles for the four matrices. Note that for each case, both directions are illustrated. Profiles 1a and 1b are based on the Avenida Grecia corridor where the load is highest in the middle segment, as is typical of services that cross a city’s central business district. Profiles 1c and 1d, based on Avenida Pajaritos, reflect the increasing or decreasing pattern of a corridor whose activity is concentrated at one end, typical of services that originate or terminate in the central business district or act as a feeder to another mass transport mode such as a Metro system.

In the 10-stop scenarios (Figures 1a and 1c) the stops are assumed to be equally spaced at 800m intervals, implying the corridors are 7.2 Km long in each direction, whereas in the 20-stop scenarios (Figures 1c and 1d) the stops are equidistant at 400m intervals so that each direction is 7.6 Km long. For both corridors we assume the operating speed of the vehicles is 25 Km/h. The key attributes of each base trip matrix are summarized in Table 3.

Table 2: Attributes of base trip matrices.

Matrix	No. of stops	Corridor	Total trips (pax/hr)	Std dev (pax/hr)	Av. no. of trips(pax/hr)	Coeff. of var.	Av. trip length (in stops)	Critical load (pax/hr)
1	10	Grecia	37,728	507.0	419.2	1.2	3.2	13,392
2	10	Pajaritos	20,546	673.7	228.3	3.0	5.1	14,119
3	20	Grecia	38,744	205.9	102.0	2.0	6.3	13,400
4	20	Pajaritos	20,453	176.0	53.8	3.3	10.6	14,101

As these data reveal, the matrices cover a variety of interesting possibilities. The values for the standard deviation and average number of trips are calculated at the O-D pair level and the ratio of these two values is the coefficient of variation. Average trip length is expressed in stops, a trip between two consecutive stops being considered as one stop long.

The possible values of the remaining parameters in the experiment are defined as explained below and summarized in Table 4.

- Travel time value as shown in the table is approximately the value used in Leiva et al. (2010) which is based on data for the city of Santiago.
- Wait time values are either 1, 2 or 3 times the travel time value since users are typically assumed to penalize waiting more than travelling.
- The base values for the operator cost components c_{L} and c_T are set at 500 Chilean pesos per bus-Km and 5,000 Chilean pesos per bus-min, the two figures derived from data for Santiago in Fernández and De Cea (2003). The values actually set for the various scenarios range as high as double the base values and as low as one-third of them but the same relative proportions are always maintained.
- Demand volume β for the different scenarios as shown in the table is also varied up to double and down to one-half.
- Dwell time values are 0.5, 1.0 and 1.5 minutes. These levels are deliberately set relatively high since preliminary testing demonstrated that with times of less than 30 seconds, express services would rarely be justified.
- Vehicle capacity values are 80, 120 and 160 pax/bus.

The parameters combine with the four base matrix cases to form a total of 972 scenarios to be optimized in our experiment.

Table 3: Range of possible parameter values for scenario generation.

Parameter	Number of options	Possible values
Base matrix	4	1, 2, 3 or 4

In-vehicle travel time ($\boldsymbol{\theta_{TT}}$)	1	15 \$/min
Wait time ($\boldsymbol{\theta_{WT}}$)	3	15, 30 or 45 \$/min
Operating cost - distance ($\boldsymbol{c_L}$)	3	250, 500 or 750 \$/bus-Km
Operating cost - time ($\boldsymbol{c_T}$)	1*	$10c_L$ \$/bus-min
Demand volume ($\boldsymbol{\beta}$)	3	0.5, 1.0, 2.0
Dwell time ($\boldsymbol{\tau}$)	3	0.5, 1.0, 1.5 min
Vehicle capacity ($\boldsymbol{b}$)	3	80, 120, 160 pax/bus

3.2 Calculation of corrected percentage savings

The corrected percentage savings for a given scenario in the experiment is defined as the percentage difference between the corrected cost (less the fixed components) of the scenario's optimal solution using express services and the corrected cost of the solution for the baseline optimal-frequency all-stop service. To obtain this percentage we must first define the minimum total trip cost as that which would obtain if every trip was made with no stops from origin to destination. This cost is given by the following formula:

$$MTTC = \theta_{TT} \sum_{w \in \mathcal{W}} T_w \sum_{l \in \mathcal{L}} t_{AS}^{w} - PTTS \tag{7}$$

As can be seen, this value consists of the travel time cost when all users take the all-stop service minus the dwell times at each stop made on their trips. The latter component is thus the potential travel time savings, whose formula is as follows:

$$PTTS = \theta_{TT} \tau \overline{Sk} T \tag{8}$$

In this equation, τ is the travel time saved by skipping a stop while $\overline{Sk}$ is the average number of skippable stops for a trip along the corridor in question. The latter term is given by

$$\overline{Sk} = \frac{\sum_{w \in \mathcal{W}} T_w \left(D(w) - O(w) - 1\right)}{T} \tag{9}$$

where $O(w)$ and $O(w)$ represent the origin and destination nodes of O-D pair w. The total corrected cost of a given solution can then be calculated as

$$CTC = TC - MTTC \tag{10}$$

If CTC* is the corrected cost of the solution of a scenario optimized by the algorithm and CTC$_{AS}$ the corrected cost of the baseline all-stop solution, the corrected percentage savings is

$$CPS = \frac{CTC^{*} - CTC_{AS}}{CTC_{AS}} \tag{11}$$

The all-stop solution corrected cost $\mathrm{CTC_{AS}}$ is derived by substituting the optimal frequency into the objective function (1), that is, setting $f_l = g_l^w = f_{AS}^{*}, \forall l \in \mathcal{L}, \forall w \in \mathcal{W}$ where f_{AS}^{*} is given by (6).

3.3 Calibration of the model

For each of the above-described 972 scenarios a set of services was designed using the algorithm developed in Section 2. The corrected percentage savings was calculated for each of these solution and the multiple regression model was then calibrated for this indicator.

To specify the model, we began by calibrating a simple model consisting of explanatory variables related to the factors already identified as significant in previous studies and cited here in Section 1: average trip length, trip concentration and system trip volume. We then gradually added new variables suspected to have explanatory power, thus improving the model at each iteration. After two iterations this procedure resulted in a model that provided a satisfactory goodness of fit ($R^2 = 86.9\%$) with no visible room for improvement. The explanatory variables retained in the calibrated model were thus the following:

- Potential travel time savings. This variable explains the same effect as average trip length but was included in its stead. Although average trip length is highly correlated with express service performance given that the longer the trip the greater the number of stops that can be skipped, it also depends on the distance between stops and thus in some cases will deliver a distorted result.
- Trip matrix coefficient of variation. As shown in Larrain (2010b), this coefficient represents the corridor trip concentration, given by the following equation:

$$CV = \frac{\sigma_w}{\overline{T_w}} \tag{12}$$

- where $\overline{T_w}$ is the average number of trips between O-D pair $w \in \mathcal{W}$ and σ_w is the standard deviation of the trips. Higher values of CV indicate that the trips are less dispersed within the matrix, implying that they are concentrated among a smaller number of O-D pairs.
- Overcrowding index. The system trip volume and especially its overcrowding level also serve as predictors of the potential of express services for a corridor (Larrain, 2013). One way of determining if a problem involves overcrowding is to calculate its optimal solution without the capacity constraint and then check whether the solution’s loads satisfy the constraint, and if not, by how much they exceed it. Taking this as the overcrowding baseline for the solution without express services, the overcrowding indicator is then formulated as

$$OI = \frac{P/b}{\sqrt{\frac{\theta_{WT}\lambda T}{c_{AS}}}} \quad (13)$$

A value of OI greater than 1 means that the capacity constraint determines the optimal frequency for the all-stop case, implying in turn that the scenario can be considered as overcrowded. More generally, the greater the OI the higher the scenario’s overcrowding level.

- Operator cost per unit of distance. This variable, incorporated into the model after the first iteration, is the c_L parameter in the operator cost formula (5).
- Dwell time. This variable was incorporated into the model after the second iteration and represents the time a vehicle takes to visit a stop, including deceleration, actual time at the stop and acceleration.

The specification of the calibrated model is as follows:

$$CPS = \alpha_0 + \alpha_1 PTTS + \alpha_2 CV + \alpha_3 CI + \alpha_4 c_L + \alpha_5 \tau \quad (14)$$

The parameter values of the calibrated model are summarized in Table 5.

Table 4: Parameter values of calibrated model.

Parameter	**Variable**	**α_i value (t statistic)**
$\boldsymbol{\alpha_0}$	-	-1,41E-01 (-18,67)
$\boldsymbol{\alpha_1}$	$PTTS$	2,58E-08 (27,15)
$\boldsymbol{\alpha_2}$	CV	8,41E-02 (38,39)
$\boldsymbol{\alpha_3}$	CI	3,24E-02 (19,24)
$\boldsymbol{\alpha_4}$	c_L	-9,35E-06 (-24,48)
$\boldsymbol{\alpha_5}$	τ	1,21E-01 (25,56)
$\boldsymbol{R^2}$	-	86.90%

4 Analysis of the experiment results

As may be observed in the specification of the model, the benefits of a given scenario are estimated by five explanatory variables: 1) potential travel time savings; 2) trip concentration; 3) system overcrowding level; 4) all-stop service operating costs; and 5) dwell time. These variables can be further decomposed to determine the effects of other factors on the problem. The five just cited are analyzed below in light of the results of the results of the experiment.

4.1 Potential travel time savings

The benefits of express services increase as potential travel time savings increase. From (8) it is evident that these savings increase in turn with increases in the following factors:

- In-vehicle time.
- Dwell time. This factor influences the regression in its own right, and does so in the same direction in both cases. The greater the dwell time, the greater is the incentive to use express services.
- The number of stops per trip. This variable is approximated in some cases by average trip length given that the two are highly correlated. Another variable typically correlated with the number of stops is the stop density since an increase in the latter implies that trips cover more stops.
- The total number of trips. Since this factor also influences the overcrowding index, the net effect will be discussed under the latter variable.

4.2 Trip concentration

As regards trip concentration, the results confirm what was observed in Larrain (2010b), namely, that the coefficient of variation satisfactorily captures this effect and is a good predictor of express service performance.

4.3 Overcrowding level

The results showed that the greater is the overcrowding, the greater are the express service's potential benefits. Recalling the overcrowding indicator (13), if we assume the maximum profile is proportional to the flows (that is, $\mathrm{P} = \mathrm{kT}$), then

$$OI = \frac{k}{b}\sqrt{\frac{c_{AS}T}{\theta_{WT}\lambda}} \qquad (15)$$

This formula implies that the benefits of express services increase with:

- Increases in the critical arc load relative to total trips.
- Decreasing vehicle capacity, which clearly translates into greater overcrowding levels.
- Decreasing wait time value multiplied by λ.
- Increasing operator costs. This effect contradicts the direct effect of operator costs as an explanatory variable in the regression; the net effect will be discussed below under operating cost.

- Decreasing number of total trips.

Since the effect of total trips on the overcrowding factor is the opposite of its effect on potential express service savings, we measure the net effect on the benefits predicted by the regression. Thus, we derive (14) with respect to total flow as follows:

$$\frac{\delta CPS}{\delta T} = \alpha_1 \theta_{TT} \tau \overline{Sk} + \alpha_3 \frac{k}{2b} \sqrt{\frac{c_{AS}}{\theta_{WT} \lambda T}} \tag{16}$$

This expression is positive for any value of T, implying in accordance with model CPS that an increase in total flow makes express services more attractive.

4.3 Operating costs

The net operating costs for all-stop service must also be calculated given the factor's multiple appearances in our formulations. The derivative of the benefits with respect to these costs is

$$\frac{\delta CPS}{\delta c_{AS}} = \alpha_3 \frac{P}{2b\sqrt{c_{AS} \theta_{WT} \lambda T}} + \alpha_4 \tag{17}$$

Given that α_4 is negative, the value of this derivative will depend on the value taken by all the other parameters. Upon evaluating the derivative for the 972 scenarios, it was found that even though in the majority of cases it was negative (i.e., an increase in operating costs reduces express service performance), there were 6 scenarios in which the derivative was positive. This implies that the effect of these costs on express service benefits depends on the specifics of a given scenario.

It is clear from (5) that operating costs for the all-stop service c_{AS} are a function of operating costs due to distance and time as well as the length and time of the service cycle. The foregoing implies that these variables will have the same effect as c_{AS} on express service benefits.

4.4 Dwell time

As noted above, dwell time affects the savings estimation both directly (with $\alpha_5 > 0$) and indirectly through the $PTTA$ variable. The two effects work in the same direction, however. Thus, longer dwell times in a given scenario are a sign that it has greater potential for express service benefits.

5 Conclusions

The results set out above give us a clear idea of the potential of express bus services as a BRT management policy under different scenarios. More particularly, the model we proposed can be used to predict the performance of such services. Needless to say, for the model's predictions to be valid in any given application the parameter values must be reasonably close to the ones used in our experiment; extrapolating to different underlying conditions risks losing validity. What we can nevertheless state with confidence based on the experiment reported here is that the attractiveness of express services increases with the following tendencies:

- Increasing dwell time,
- Increasing number of stops per trip or average trip length,
- Increasing number of trips on the system,
- Increasing travel time,
- Increasing concentration of trips on a limited number of origin-destination pairs,
- Increasing critical arc load as a proportion of the load for all trips,
- Decreasing vehicle capacity,
- Decreasing value of wait time.

One effect our proposed methodology does not directly take into account is the number of transfers made by users of express services. It can nevertheless be argued that lower transfer costs (obtainable in practice by improving transfer conditions) would be an incentive for users to take express services. An analysis that included this effect would require a frequency optimization model that generated reliable solutions in the presence of overcrowding. The authors of this study are currently working on a bi-level approach to the solution of this problem that would conclusively establish the importance of including transfers in express service design.

An interesting line of inquiry for future research would be to develop the theoretical basis of our results using analytical expressions for the optimal social costs as a function of the parameters chosen here. In an initial effort a simplified corridor could be employed with a model, also simplified, of user behaviour, and limiting the type of express service to formats that facilitate optimization.

Acknowledgements: The authors are grateful for funding provided under FONDECYT project 11140443. This research also benefited from the support of the Bus Rapid Transit Centre of Excellence, funded by the Volvo Research and Educational Foundations (VREF) and the Centro de Desarrollo Urbano Sustentable (CEDEUS), Conicyt/Fondap/15110020.

References

Chiraphadhanakul, V. and C. Barnhart (2013). Incremental bus service design: combining limited-stop and local bus services. **Public Transport**, 5(1-2), 53-78.

Desaulniers, G. and M. Hickman (2007). Public Transit. In: Barnhart, C. and Laporte, G. (Eds.). Handbooks in Operations Research and Management Science, Vol. 14, Transportation. Amsterdam, North-Holland, 69-128.

El-Geneidy, A.M. and J. Surprenant-Legault (2010). Limited-Stop Bus Service: An Evaluation of an Implementation Strategy. Public Transport, 2(4), 291-306.

Ercolano, J.M. (1984). Limited-Stop Bus Operations: An Evaluation. Transportation Research Record, 994, 24–29.

Fernández, J.E. and J. De Cea (2003). Análisis Modernización de Transporte Público, VI Etapa. Technical report. SECTRA, Santiago.

Furth, P.G. (1986). Zonal Route Design for Transit Corridors. Transportation Science, 20(1), 1-12.

Hidalgo, D. and J.C. Munoz (2014). A Review of Technological Improvements in Bus Rapid Transit BRT and Buses with High Level of Service BHLS. Public Transport, 6(3), 185-213.

Jansson, J. O. (1980). A Simple Bus Line Model for Optimization of Service Frequency and Bus Size. Journal of Transport Economics and Policy, 14, 53-80.

Jordan, W.C. and M.A. Turnquist (1979). Zone Scheduling of Bus Routes to Improve Service Reliability. Transportation Science, 13(3), 242–268.

Larrain, H., J. C. Muñoz (2008). Public Transit Corridor Assignment Assuming Congestion Due To Passenger Boarding And Alighting. Networks & Spatial Economics, 3(8), 241-256.

Larrain, H., R. Giesen and J. C. Muñoz (2010a). Choosing the Right Express Services for Bus Corridor with Capacity Restrictions. Transportation Research Record, 2197, 63-70.

Larrain, H., Muñoz, J.C., Giesen, R. (2010b) The effect of OD trip dispersion versus concentration in express service design. Working paper N°05, BRT Centre of Excellence.

Larrain, H. (2013). Diseño de Servicios Expresos para Buses. PhD thesis. Pontificia Universidad Católica de Chile, Chile.

Levinson, H. S., S. Zimmerman, J. Clinger and C. S. Rutherford (2002). Bus Rapid Transit: An Overview. Journal of Public Transportation, 5(2).

Leiva, C., J. C. Muñoz, R. Giesen, and H. Larrain. (2010) Design of Limited-Stop for an Urban Bus Corridor with Capacity Constraints. Transportation Research B, 44(10), 1186-1201.

Mohring, H. (1972). Optimization and Scale Economies in Urban Bus Transportation. American Economic Review, 62, 591-604.

Silverman, N.C. (1998). Limited-Stop Bus Service at New York City Transit. Journal of Transportation Engineering, 124(6), 503-509.

Sun, C., W. Zhou, and Y. Wang (2008). Scheduling Combination and Headway Optimization of Bus Rapid Transit. Journal of Transportation Systems Engineering and Information Technology, 8(5), 61-67.

Tétreault, P. and A. El-Geneidy (2010). Estimating Bus Run Times for New Limited-Stop Service Using Archived AVL and APC Data. Transportation Research Part A, 44(6), 390–402.