Holding times to maintain quasi-regular headways to reduce real-time bus bunching[1]

Citlali M. Olvera-Toscano
Universidad Autónoma de Nuevo León (UANL)
Graduate Program in Systems Engineering, San Nicolás de los Garza, Mexico
E-mail: *citlali.olverats@uanl.edu.mx*

Yasmín Á. Ríos-Solís
Tecnologico de Monterrey, Engineering and Sciences Department
Monterrey, Mexico
E-mail: *yasmin.riossolis@tec.mx*

Roger Z. Ríos-Mercado
Universidad Autónoma de Nuevo León (UANL)
Graduate Program in Systems Engineering, San Nicolás de los Garza, Mexico
E-mail: *roger.rios@uanl.edu.mx*

Romeo Sánchez Nigenda
Universidad Autónoma de Nuevo León (UANL)
Graduate Program in Systems Engineering, San Nicolás de los Garza, Mexico
E-mail: *romeo.sanchezng@uanl.edu.mx*

September 2021

[1]Preprint PISIS-RR21-03, Graduate Program in Systems Engineering, UANL, San Nicolás de los Garza, Mexico, September 2021

Abstract

Real-time control strategies palliate with the day's dynamics in bus rapid transit systems. In this work, we focus on a bus bunching problem that minimizes the number of buses of the same line cruising head-to-tail or arriving at a stop simultaneously by using bus holding times at the stops. For this, we propose a new mathematical model with quadratic constraints, whose objective function minimizes the penalties caused by buses that are bunching. Experimental results on a simulation of a bus rapid transit system in Monterrey, Mexico, show the efficiency of our approach. The results show a bus bunching reduction of 45% compared to the case without optimization. Moreover, in some scenarios the passenger waiting times are reduced by 30%.

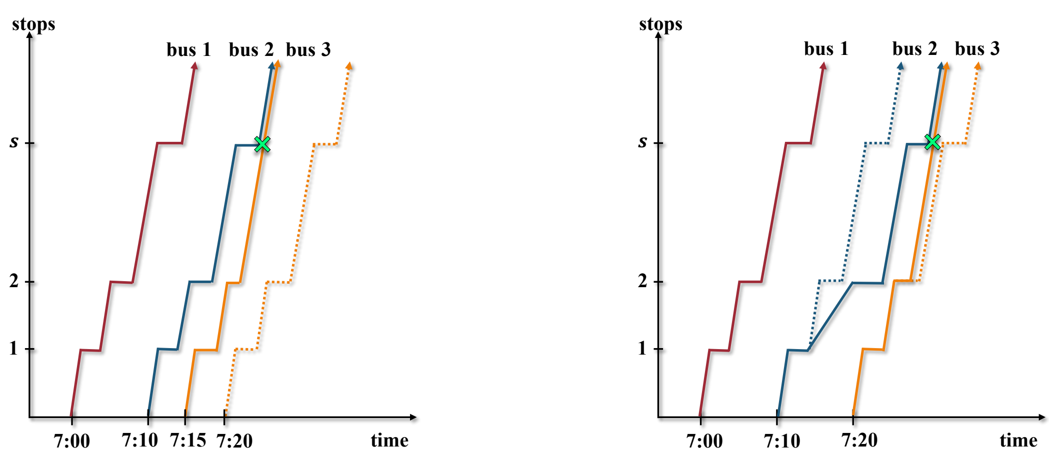
Keywords: Bus bunching; Real-time; Holding time; Transit operations.

1 Introduction

Almost any city encourages public transport to mobilize its population more sustainably. Nevertheless, public transport systems often lack high-quality service, making them unattractive to users with a private vehicle [35, 20]. Public transportation companies aim to reduce variability in the user's waiting times at the stops to increase customer service satisfaction and attract more customers since quality is positively correlated to more reliable systems [4, 27]. For this, technology and infrastructure in transportation services enable better communication with the bus drivers and extensive data collection such as user flows, bus occupancy, geospatial location of the buses, or road reports. This information provides real-time feedback to the bus control system that aims to improve the users' satisfaction [35, 4, 10].

Once a public transportation company establishes the system planning, that is, the buses' departure times are set for each stop with a specific frequency, the company uses real-time control strategies to palliate with the day's dynamics such as traffic, passenger flow, weather, or accidents. As Wang and Sun [36] and Ceder [4] mention, a slight frequency deviation can cause many buses to be delayed on their schedule.

In this work, we focus on a *bus bunching problem* (BBP)[9, 21, 29, 36], where we seek to determine a set of bus holding times at some stops to minimize the number of buses of the same line cruising head-to-tail or arriving at a bus stop simultaneously. Bus bunching implies frequency variability that affects the user's waiting times at the stops and traveling times. Also, bus bunching involves unbalanced occupancy rates since many waiting users choose to board the first bus that arrives at the stop, especially when the information on the arrival of the next bus is unknown, as in most cities in developing countries.



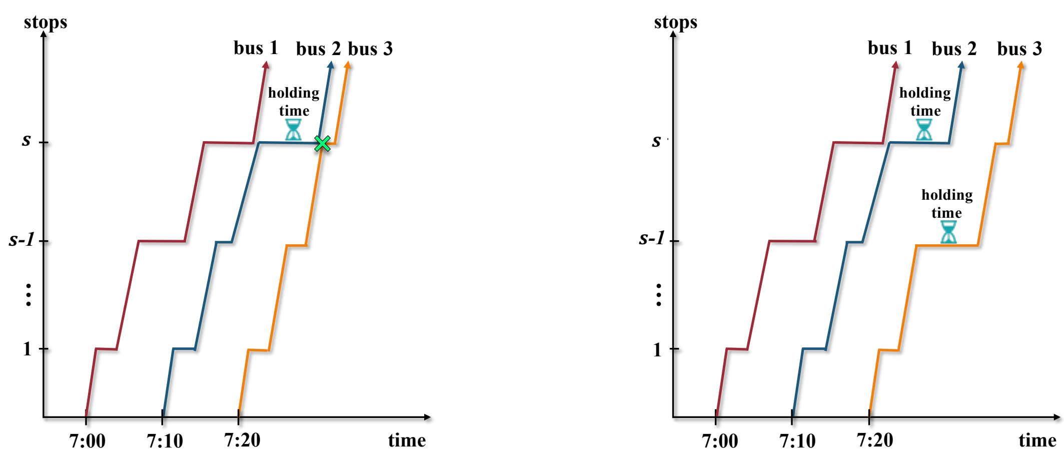
(a) Early departure time.

(b) Delayed bus by traffic.

Figure 1: Causes of bus bunching (modified from Ceder [4]).

Figure 1 shows the leading causes of bus bunching for a single bus line. Time is plotted on the horizontal axis, while bus stops are on the vertical axis. There are three trips (red, blue, and yellow) with an established departure time of 7:00, 7:10, and 7:20, respectively. Each plot shows the time-space trajectories of the three trips. In Figure 1a, the first two trips (red and blue) behave as planned: at any time and space, the time difference of the trips (known as headway) is equal to ten minutes. Nevertheless, the bus departure of the third trip (solid yellow trajectory) is five minutes early (dotted yellow trajectory is the ideal one). This action causes fewer users than expected at stop 1, so the dwelling time of this bus is shorter than expected, and eventually, it bunches with the second bus. In Figure 1b, the second bus (blue) finds more congestion than usual between stops 1 and 2. Thus, there are more persons at stop 1, implying a larger dwelling time. The blue line takes, at stop 2, some passengers that should have taken the yellow trip—eventually, the blue and the yellow trip bunch. A similar case happens when more people than expected arrive at a stop: a concert or sports event.

Our objective is to reduce the bus bunching by maintaining quasi-regular headways between each pair of buses to reduce users' waiting times at the stops and their overall traveling times. For this, we use a control strategy consisting of holding the buses for some minutes at some stations after the boarding and alighting processes have been completed.



(a) Applying holding time to a bus.

(b) Applying holding time to multiple buses.

Figure 2: Decreasing bus bunching with the holding time strategy.

Figure 2 shows how the holding bus strategy helps to reestablish regular headways. In Figure 2a, there is a bunching caused at stop s by buses 2 and 3. We could apply a holding time to bus 2 at station s to avoid this bunching. However, retaining a single bus (as some local optimization strategies do) is not enough since bus 3 will bunch up eventually with bus 2 at a future stop. Thus a global tactic, applying holding times to several buses (see Figure 2b), is a better real-time strategy.

The proposed solution approach to this BPP has two main processes, as shown in Figure 3. One of them is an optimization stage that determines the holding times of each bus at each stop by considering the current information along the bus line and by predicting the future events in the line: persons at each bus stop, dwelling and alighting times, and the capacity of each one of the buses. This stage is modeled as a quadratically constrained integer linear programming that yields the holding times for the buses, which are immediately communicated to the drivers to recover quasi-regular headways between them. This model is one of the main contributions of this work. While most related studies deal with a non-linear objective function that minimizes the user's total waiting time at the optimization stage (see the overview by Ibarra-Rojas et al. [23]), our BBP methodology uses a linear objective function that speeds up its computational time. One of main model advantages is that it includes the forecasting of the departure, boarding, and alighting times for each bus at each future stop more precisely than previous works [9, 21]. Only a few approaches consider the bus capacity every time the optimization model computes holding times during the evolution of the simulation process, as we do in this study. A significant feature of our model is that it does not use origin-destination matrices as most of the other approaches [8, 21, 29, 36, 39], which affects their efficiency. Instead, we use the user's arrival rate and the proportions of users that alight at stops.

The other process of our methodology (Figure 3) is collecting data in real-time: the position of the buses, the actual number of users waiting at stops, and the actual number of users on each bus. This information is the input of the optimization stage. These two processes are periodically applied one after the other along the working day to maintain quasi-regular headways between each pair of buses. We developed a discrete event simulation program, representing a bus rapid transit (BRT) system, to simulate real-data collection and stress the transit system under different parameters. Thus, we can retrieve simulated real-time data from the system to validate the efficiency of our mathematical model. This simulation takes as input the holding times of the optimization stage. Then, it evolves along the time, and when asked, it returns its actual state, which is the input of the optimization stage.

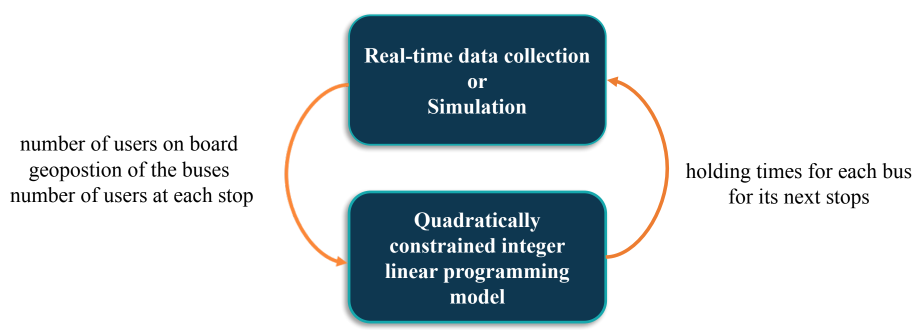


Figure 3: Framework for the real-time data retrieving and the mathematical model.

Maintaining quasi-regular headways in a transit line is different from sticking to the buses' planned departure times at the stops. Quasi-regular headways imply a reliable service for users since the variance in the frequency of the buses is small at each bus stop. This tactic is particularly helpful in cities where the users do not know the exact scheduling of the buses since they only have an expected frequency estimation as is the case of systems in developing countries [21].

This paper is organized as follows. In Section 2, we present a related literature review. Section 3 describes the system characteristics and the description of the mathematical model, as well as the discrete event simulation that represents a bus rapid transit system. We validate the efficiency of our BBP methodology with a case study from the city of Monterrey in Mexico in Section 4. Finally, Section 5 presents the conclusions of this work.

2 Literature review

Significant efforts have been made to address the BBP with real-time monitoring strategies, categorized in *inter-station* and *station control* [23]. Inter-station control strategies are those where decisions are made at some bus line stops such as holding times [36, 6, 39], skipping stops [31, 15, 5], or boarding time limits [9, 2, 8]. station control strategies are bus speed regulation [18, 7, 19] or traffic signal priority [13, 26, 25].

Among all the BBP strategies, holding time is the most used and the one that users resent the least. Table 1 shows the studies related to the holding time strategy (we do not include hybrid strategies). The first column presents the reference. The second column indicates whether the vehicle capacity is considered or not. The third column indicates the different types of control points that the system considers, such as single preset control point (SPCP), multiple preset control points (MPCP), or multiple control points (MCP) defined by the solution strategy. Column "Alighting" is how the users leave the transport system in the solution approach: a proportion of the onboard passengers or determined by an origin-destination (OD) matrix. The fifth and sixth columns indicate the considered objective function and model type, respectively (QP is for quadratic MILP, MIQCP is for quadratically constrained integer linear programming). The solution approach is in the "Method" column (B&B is for branch-and-bound), while the "Sim" one indicates whether the approach is analyzed through the use of any simulation. Finally, the last column reports if a case study was tested.

The last line of Table 1 considers the approach we propose. Our BPP approach searches for headway regularity which we show later that implies reducing waiting times. This search for headway regularity allows us to have a linear objective function that reduces computational times compared to other approaches that directly optimize waiting times with a quadratic function. We also consider that buses have a capacity, and the users must wait for the next bus if this capacity is reached. Our optimization process determines the stops where the buses must hold and the holding times. An enormous gain is obtained by considering that

Author	Cap	Control points	Alighting	Objetive Func.	Model type	Method	Sim	Case study
Eberlein et al. [12]	no	SPCP	proportional onboard passengers	minimize waiting time	QP	local search	yes	yes
Hickman [22]	no	SPCP (one bus)	OD matrix	minimize headway variance	stoch. QP	gradient search	no	no
Sun and Hickman [32]	no	MPCP	proportional onboard passengers	minimize waiting time	QP	heuristic	no	no
Xuan et al. [39]	no	MPCP	OD matrix	schedule adherence and headway regularity	QP	local gradient search	no	no
Delgado et al. [9]	yes	MPCP	OD matrix	minimize waiting time	QP	reduced gradient	yes	no
Hernández-Landa et al. [21]	yes	MCP	OD matrix	headway regularity	MILP	B&B	yes	yes
Sánchez-Martínez et al. [29]	yes	MCP	OD matrix	minimize waiting time	QP	iterative quadratic approximation	yes	no
Wang and Sun [36]	no	MCP	OD matrix	reduce the headway mean and variance	QP	multi-agent system	yes	no
Our approach	yes	MCP	proportional onboard passengers	headway regularity	MIQCP	B&B	yes	yes

Table 1: Literature review of the holding time strategy.

passengers alight a bus via a probabilistic approach since an OD matrix is rarely updated and reliable.

Argote-Cabanero et al. [1] employ a combination of dynamic holding times and en-route driver guidance to improve schedule adherence. This method is analytically evaluated with simulations to improve reliability. As in our work, their simulation reflects a real BRT. Hall et al. [17] develop analytical models to obtain the optimal holding times and waiting times at transfer stations. Their instances are based on actual data from the Los Angeles transit network. Daganzo [6] propose an adaptive control scheme that dynamically determines holding times at the stops. The proposed scheme's objective is to provide quasi-regular headways while maintaining a commercial speed. The method proves to be effective when minor disturbances arise.

More recently, Gkiotsalitis and Cats [16] employ a periodic holding time control method where holding times of all running trips are computed simultaneously within each optimization period. They model the BBP as a discrete non-linear optimization problem. They do not consider additional user's arrivals when the bus is waiting at a stop neither the capacity of buses since all users can board the first bus that arrives, even if this implies longer headways.

He et al. [20] present a dynamic target-headway-based holding strategy validated by a numerical experiment. They observe that the average total waiting time increases at the beginning of their experiment but decreases along the time, which is the desired behaviour. Delgado et al. [9] incorporate two strategies: holding time and boarding limits. They propose a mathematical programming model that minimizes total delays with a quadratic function evaluated in a simulation environment. Nevertheless, their simulation considers very short headways, computing holding times when a bus reaches a stop. Although they obtain the value of the holding times for all the buses in their following stops, they only use the holding time value for the current stop discarding the other values. Wang and Sun [36] incorporate global coordination and long-term operation in holding time with a multi-agent deep reinforcement learning framework. However, they do not consider the bus capacity. Sánchez-Martínez et al. [29] present a quadratic model to compute passengers' average waiting times costs and separately obtain the holding times. Their simulation only considers ten buses and equal arrival rates for all origin-destination pairs.

In this paper, we integrate simulation and optimization to reduce the number of bus bunching and indirectly, reduce the user's total waiting times. On the one hand, we develop a discrete event simulation of a BRT system. On the other hand, c The mathematical model has quadratic constraints, but its objective function is linear, speeding up its computational time.

3 The BBP methodology

Our BBP solution approach has two main processes (see Figure 3). In this section, we describe our bus rapid transit (BRT) system and introduce its notation. Then, we present the mathematical model of the optimization stage that determines the holding times of each bus. Finally, we present the simulation stage that represents the real-time data collection of a BRT case study.

3.1 The bus rapid transit description

Consider a circular bus line corridor with a set S of stops plus the depot denoted as 0, as represented in Figure 4, operated by a high-frequency bus service. A set of B of buses with capacity C_b for each $b \in B$ and average speed of V constitute the bus service. All buses start their trip at the depot following a departure frequency of F minutes. Each bus $b \in B$ sequentially visits all the stops in the corridor. The buses are numbered in ascending order in the corridor; bus $b+1$ leaves the deposit after bus b, for $b \in B \setminus \{|B|-1\}$. Notice that bus passing is not allowed. When a bus b arrives at the last stop, all bus users must alight, and the bus returns to the depot (no deadheads, since it is a circular corridor). The travel time between a pair of stops $(s-1, s)$ is denoted as d^s, for $s \in S$. The limit case d^1 is the distance between stop 1 and the depot.

The users arrive at each stop $s \in S$ following a Poisson distribution rate λ^s. The boarding and alighting times per user are $\overline{U}$ and $\underline{U}$, respectively. The total time taken by the bus doors for opening and closing is G. Similarly to Eberlein et al. [12], our work assumes that the number of users alighting from bus $b \in B$ at stop $s \in S$ is a proportion γ^s of the number of onboard users, a crucial point since we avoid using origin-destination matrices that make heavier models. Additionally, the holding time assigned to a bus at a stop must not exceed a number H of minutes per station to avoid the despair of the onboard users. This holding limit is an essential aspect of our model and its solution efficiency.

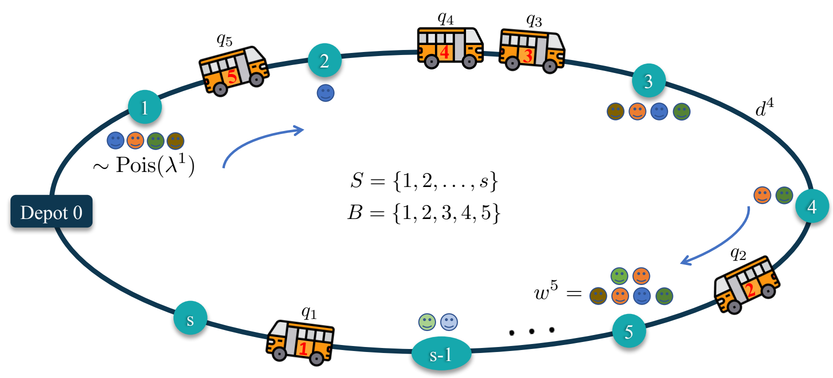


Figure 4: Circular bus line corridor with s stops (green circles) and five buses.

Time t_0 is when the real data is collected to compute the required decisions to recover a quasi-regular frequency. We collect the following data for each $b \in B$: the distance between bus b and its following stops denoted as m_b^s, the number of onboard users on bus b is q_b, and l_b is the last station visited by bus b. Note that if bus b is at station s at time t_0, then $l_b = s$. Also, at time t_0, we collect the number of users waiting at station s denoted by w^s for every station $s \in S$.

3.2 The BBP optimization stage

The main variables to formulate the quadratically constrained integer linear programming of the BBP methodology are h_b^s that indicate the holding time for bus $b \in B$ for all the future stop $s = l_b + 1, \ldots, |S|$ that the bus must visit. With the position of the buses at time t_0 and the holding time variables, we can derive the rest of the variables of each bus $b \in B$ at stop $s = l_b + 1, \ldots, |S|$. Let t_b^s be the variables that represent the departure time of bus b from stop s. With these last variables, we can compute the number of users who want to board bus b at stop s, the number of users who indeed boarded the bus, and the number of persons that alight that bus, denoted by z_b^s, x_b^s, and y_b^s, respectively, for $b \in B$ and $s = l_b + 1, \ldots, |S|$.

The headway between the bus pair $(b, b+1)$ is defined by the difference of its departure times at a stop s: $t_b^s - t_{b+1}^s$, for $s \in S$. Figure 5 shows the headway values between a pair of buses at any fixed stop in its horizontal axis, while the vertical axis represents the headway penalty function. Since the buses cannot overtake each other, then this difference is positive. Let κ be the small headway deviation percentage from the ideal frequency F. Thus, there is a bus bunching between b and $b+1$ if their headway is between $[0, (1-\kappa)F]$. Consequently, b and $b+1$ have a quasi-regular headway if it is between $[(1-\kappa)F, (1+\kappa)F]$. The penalty is a linear function proportional to the headway between a pair of buses and is denoted by p_b^s, for $b \in B$ and $s \in S$. When bus pair $(b, b+1)$ has a headway larger than $(1+\kappa)F$, they incur a significant penalty:

$$p_b^s \geq t_b^s - t_{b+1}^s + (1-\kappa)F, \qquad b \in B, s = l_b + 1, \tag{1}$$

$$p_b^s \geq t_{b+1}^s - t_b^s + (1+\kappa)F, \qquad b \in B, s = l_b + 1. \tag{2}$$

Constraints (1) and (2) define the penalty value of bus b at stop $l_b + 1$. On the one hand, constraints (1) are activated when buses b and $b+1$ are bunching. On the other hand, constraints (2) are used when the buses are farther apart than allowed. Note that p_b^s cannot have negative values since it is defined as a positive variable.

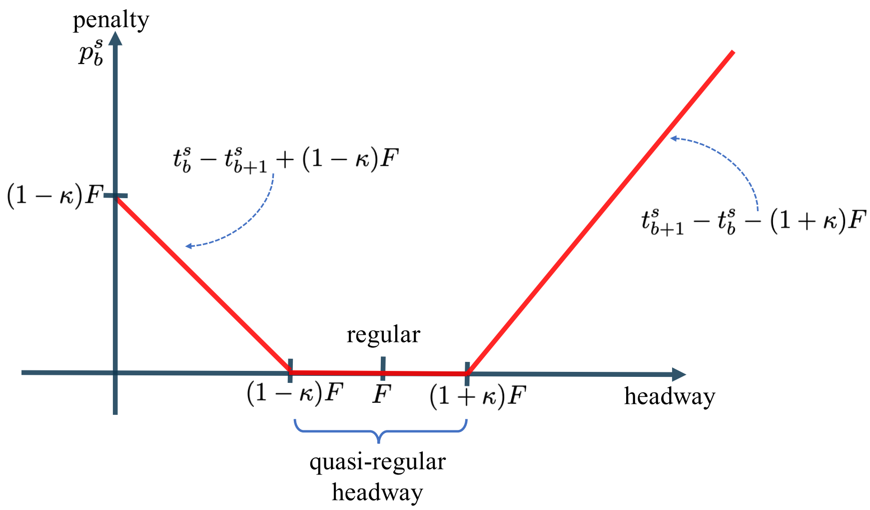


Figure 5: The penalty function used by the objective function (3) between a bus pair at a certain stop.

In this way, each pair of buses complement each other in the objective function (3) that minimizes the bunching penalties for each pair of buses, which results in maintaining quasi-regular headways between the buses:

$$\min \sum_{b \in B} \sum_{s=l_b+1}^{|S|-1} p_b^s. \tag{3}$$

By minimizing the number of bus bunching, we also minimize the user's waiting times, as shown in the experimental results.

For establishing the departure times of the buses at each future station that they will visit after time t_0, there are two cases. For handling the first case, constraints (4), define the departing time of bus b at stop $l_b + 1$ as the time t_0 plus the remaining time to arrive at stop $l_b + 1$, the boarding and alighting times, its holding time, and the opening and closing of the bus doors. For the second case, constraints (5) define the departure time of bus b from stop $s = l_b + 2, \dots, |S|$. Constraints (6) do not allow a bus to leave a stop earlier than its predecessor, thus ensuring that passing between buses is impossible. Constraints (7) ensure that the holding times do not exceed the maximum allowed value. These last constraints were also included in the model by Hernández-Landa et al. [21].

$$\begin{aligned}
t_b^s &= t_0 + \frac{m_b^s}{V} + \overline{U} x_b^s + \underline{U} y_b^s + h_b^s + G, \qquad b \in B, s = l_b + 1, && (4)\\
t_b^s &= t_b^{s-1} + \frac{d^s}{V} + \overline{U} x_b^s + \underline{U} y_b^s + h_b^s + G, \qquad b \in B, s = l_b + 2, \dots, |S|, && (5)\\
t_b^s &\geq t_{b-1}^s, \qquad b \in B, s = l_b + 1, && (6)\\
h_b^s &\leq H, \qquad b \in B, s = l_b + 1. && (7)
\end{aligned}$$

The number of users at stop s who are waiting to board bus b is established by constraints (8). They consider the persons w^s that are already at the stop at time t_0 plus the persons that arrive with a Poisson rate during the time that takes the bus to arrive at this stop s:

$$z_b^s = w^s + \lambda^s \left(\frac{m_b^s}{V} \right), \qquad b \in B, s = l_b + 1, \dots, |S|. \tag{8}$$

Note that not all these users z_b^s will be able to board bus b at stop s since we are considering that buses have a finite capacity.

Constraints (9) determine the number of alighting users from bus b at stop s. They consider the onboard passengers and the users that already alighted in previous stops $s' < s$, multiplied by the estimated proportion γ^s of passengers alighting at this stop, for $s \in S$.

$$y_b^s = \gamma^s \left(q_b + \sum_{s'=l_b+1}^{s-1} x_b^{s'} y_b^{s'} \right), \qquad b \in B, s = l_b + 1, \dots, |S|. \tag{9}$$

The boarding users at stop s is determined with the following constraint (10), for $b \in B, s = l_b+1, \dots, |S|$:

$$x_b^s = \min \left\{ C_b - q_b - \sum_{s'=l_b+1}^{s} x_b^{s'} + \sum_{s'=l_b+1}^{s} y_b^{s'}, z_b^s - \sum_{b'=0}^{b-1} x_{b'}^s \right\}. \tag{10}$$

The right side of this minimum considers the case where the number of users waiting at s is less than the bus capacity. Thus all users board except the ones that already boarded a previous bus. The left side of the minimum limits the boarding users to the bus current capacity and subtracts the users who boarded previous buses. Constraints (10) are not linear, but a classical linearization [37] for minimum cannot be applied in this case because the objective function has positive and negative coefficients. Thus, we introduce two indicator binary variables for $b \in B$ and $s = l_b + 1, \dots, |S|$. The first one, n_b^s, is equal to 1 if the number of users waiting for bus b at stop s is less than the actual bus capacity, and 0 otherwise. The second one, r_b^s, equals 1 if the available bus capacity is less than the number of users waiting for bus b at stop s, and is

0 otherwise. With these two binary sets of variables, we replace equation (10) with linear constraints (11) and (12), for $b \in B$ and $s = l_b + 1, \ldots, |S|$:

$$
\begin{aligned}
1 &= n_b^s + r_b^s, && (11)\\
x_b^s &= n_b^s \left(z_b^s - \sum_{b'=0}^{b-1} x_{b'}^s \right) + r_b^s \left(C - q_b - \sum_{s'=l_b+1}^{s} x_b^{s'} + \sum_{s'=l_b+1}^{s} y_b^{s'} \right). && (12)
\end{aligned}
$$

Constraints (11) indicate that only one case is possible. Either all the users waiting at stop s are going to board bus b, implying $n_b^s = 1$, or some of them will wait for the next bus $b+1$; thus $r_b^s = 1$. By adding constraints (12) we explicitly choose between these two options and assign the correct value to the boarding variable x_b^s.

To reinforce the mathematical model, we add two families of valid inequalities. Constraints (13) and (14) bound the boarding passengers variable x_b^s by either the passengers at the stop or its remaining capacity, respectively, for $b \in B$ and $s = l_b + 1, \ldots, |S|$. These constraints are not necessary for the model to be correct, but they strengthen the convex hull of the discrete solution space, which reduces the resolution time of the mathematical model [38].

$$
\begin{aligned}
x_b^s &\le z_b^s - \sum_{b'=0}^{b-1} x_{b'}^s, \qquad b \in B, s = l_b + 1, \ldots, |S|, && (13)\\
x_b^s &\le C_b - q_b - \sum_{s'=l_b+1}^{s} x_b^{s'} + \sum_{s'=l_b+1}^{s} y_b^{s'}, b \in B, s = l_b + 1, \ldots, |S|. && (14)
\end{aligned}
$$

To summarize, the mathematical model that determines the best holding times for the buses at their future stops at time t_0 is named H-BPP(t_0):

$$
\begin{aligned}
\min \quad & (3),\\
\text{s.t.} \quad & (1), (2),\\
& (4) - (9)\\
& (11), (12),\\
& (13), (14), \qquad \{\text{valid inequalities}\}\\
& h_b^s, x_b^s, t_b^s, z_b^s, y_b^s \in \mathrm{R}^+,\\
& n_b^s, r_b^s \in \{0, 1\}.
\end{aligned}
$$

In the H-BBP(t_0) model, the forecast of the departure times and the boarding and alighting times play an essential role for each bus at each future stop. These approximations are one of the critical points of our study since they are much more precise than the ones made by other related approaches [21, 9].

3.3 The simulation of the real-time collection stage of the BBP

Our BBP methodology requires real-time data collection. We can obtain this data by using global positioning or automatic vehicle location systems for each bus in the BRT system. However, this kind of data is not always available, so we developed a discrete-event microsimulation to mimic a real-time situation described in Section 3.1. Besides, with our simulation, we can analyze the impact of the optimization model under different scenarios.

Our discrete-event simulation model is stochastic (with random components on the travel times and the arrival of the users) but is not dynamic since time is not variable [24, 30]. Indeed, we simulate a single rush-hour period.

The system state of our microsimulation is composed of four components that contain enough information to describe the evolution of the transportation system over time: the activation of a bus, the user's generation, the movement of the buses that consider the boarding and alighting users, and the holding times obtained by the optimization stage.

The `ActiveBus` function ensures that every F minutes, a new bus leaves the depot, recording all its information at every step of the simulation: the departure times, the number of persons on board, or alighting, travel times, and holding times.

At each step of the simulation and for each stop s, the `UserGeneration`(s) function randomly generates users following a Poisson distribution with mean λ^s, for each $s \in S$. Each generated user at s is recorded together with its waiting time to catch the first bus and its traveling time on board. In this manner, at the end of the simulation, we can compute the total waiting time of the users.

Algorithm 1 MoveBus(s, b)

let $s' = l_b + 1$
while bus is not at stop s **do**
  advance b up to s' during m_b^s/V of time
  at stop s', hold bus b of $\overline{U}\bar{x}_b^{s'} + \underline{U}\bar{y}_b^{s'} + h_b^{s'} + G$ time
  $s' = s' + 1$
end while

The `MoveBus`(s, b) function, described by Algorithm 1, describes how the buses move in the BRT system. During each simulation step, a bus b may be advancing or stopped at a stop. A bus dwells at a stop $s' < s$ until the users board and alight $\bar{x}_b^{s'} + \bar{y}_b^{s'}$, plus the holding time h_b^s established by the H-BBP(t_0) model, plus the time needed by the doors to open and close. This addition reflects the case where boarding and alighting take place through the same door. Note that $\bar{x}_b^{s'}$ and $\bar{y}_b^{s'}$ are not variables; they are values computed with the actual users at the stops and the actual capacity of the bus. Alighting users $\bar{y}_b^{s'}$ is a proportion $\gamma^{s'}$ of the onboard users. The boarding process follows a First-in-First-out (FIFO) rule, so the first user at the stop is the first to board the bus, provided that the bus has available capacity. Otherwise, the user waits until the next bus arrives with available capacity. Therefore, $\bar{x}_b^{s'}$ is the minimum value between the available capacity of the bus b and the users waiting at a stop. The holding times are obtained from the H-BBP(t_0) model presented in Section 3.

At every step of the simulation, we keep track of the bus capacities and their location, the onboard users, and the ones at the stops. Therefore, different measures of effectiveness (MOE) may be calculated. The MOEs reflecting the system performance statistics are the total waiting time of the users to board a bus, the total travel time of the users, and the number of bus bunching events.

At the initial state of the simulation, there are only $B = \lfloor \frac{S}{2} \rfloor$ active buses in the corridor, uniformly distributed through the stops. The total execution time of the simulation represents a high demand period of $T_{\max}$ minutes. Ω is the frequency at which the holding times are computed by the H-BBP(t_0) model where t_0 is distributed evenly in the interval $[10\%T_{\max}, 90\%T_{\max}]$ (the planning period starts at time 0). The first 10% of the simulation time is the *warm-up time* to reach a steady point, while the last 10% is mainly for the users to alight the buses, so there is no much interest in computing holding times anymore at the end of the period.

4 Experimental results

In this section, we test a set of experiments on a case study with various scenarios to verify the effectiveness of our BBP strategy to mitigate the bus bunching effect and reduce the user's waiting times.

The system's current state at time t_0 is the input of the mathematical model H-BBP(t_0) receives. This model is coded in C++ and solved using the commercial solver Gurobi 8.11. Since we have a linear objective function, valid inequalities, and we avoid origin-destination matrices, our formulation can be solved in less than 15 seconds for most of the instances. In this manner, the rapidly obtained holding times are input for the discrete event simulation coded in R 3.6.0 every Ω minutes along the planning period.

The experiments were carried out on a computer with macOS Catalina 10.15.5 with an Octa-core of 3 GHz Intel Xeon E5 processor and 16GB 1866MHz DDR3 of RAM. All the instances and the code of the mathematical model are available at `https://doi.org/10.6084/m9.figshare.16688770.v1`.

4.1 Case study: Ecovía in Monterrey, Mexico

The case study analyzed in this work contemplates the morning rush-hours (6:00-8:00) of the Ecovía, a BRT corridor that provides service in Monterrey city and its metropolitan area in Mexico. The corridor has 40 stations, including two terminal stops (Lincoln and Valle Soleado) located at opposite ends of the line, as schemed by the red line in Figure 6. We consider the BRT Ecovía from Lincoln to Valle Soleado, which length is around 30.2 km. Notice that it intersects two city underground rapid transit rail system lines: Line 1 at

Mitras station and Line 2 at Regina station. Mitras is the most critical stop since this bus/subway station has higher arrival and alighting rates. Although the Ecovía BRT has three types of buses with different capacities, 80% of the buses are of the same class, so our experiments are performed with only the type of buses with a maximum capacity of 80 persons.

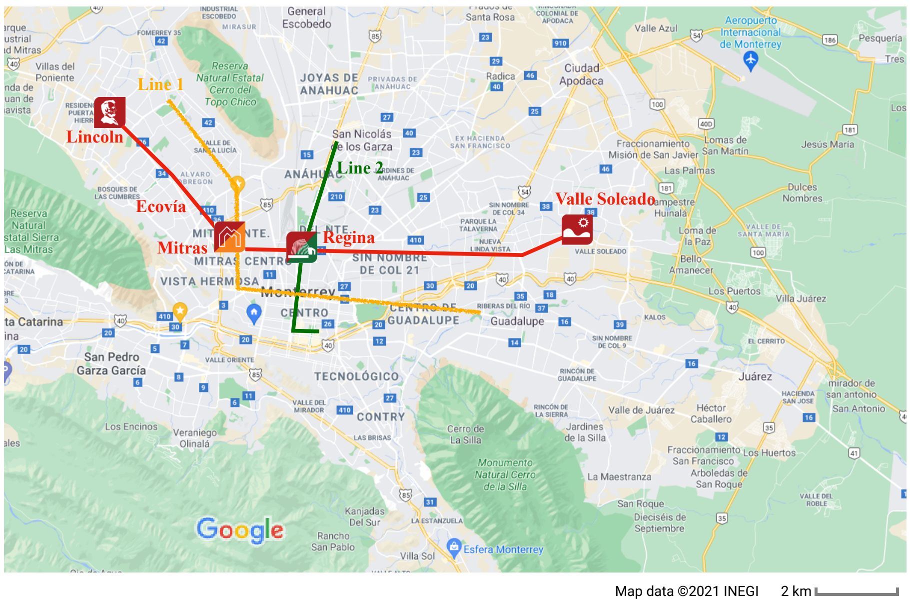


Figure 6: Map of Monterrey, Mexico. Bus rapid transit Ecovía (red) intersects Line 1 (yellow) and Line 2 (green) of the underground rapid transit rail system. Adapted with permission of the author from [Ecovía Monterrey], from Google, n.d., `https://www.google.com/maps/@25.6983801,-100.3365817,12.41z`. All rights reserved 2021 by Google.

The Ecovía simulation period is of $T_{\max} = 120$ minutes with a frequency F of two minutes during the rush hour. Thus, there could be more than 60 buses at the same time in the corridor.

We have statistically determined parameter vectors $(\lambda^s)_{s\in S}$ and $(\gamma^s)_{s\in S}$ used in the simulation and the H-BBP(t_0) model through a field survey. At each stop, the users' arrival process λ^s, $s \in S$, is a Poisson distribution with around eight passengers per minute rates at Mitras or Lincoln stations, while as few as 0.12 passengers per minute for the final stations close to Valle Soleado. For the H-BBP(t_0) model, we use the expected value of these distributions. We have also determined the proportion of users alighting at each station. The first stops have a small proportion of $\gamma^s = 1\%$, while in Mitras it is $\gamma^{\text{Mitras}} = 75\%$. Also, from our field survey, we fixed the boarding and alighting time per person in $\bar{U} = \underline{U} = 2$ seconds, similar to what was reported in Fernández et al. [14], Tirachini [34]. These times already consider a congested rush-hour system with passenger friction since users board and alight by the same door. When a bus arrives at a station, the process of opening plus closing its doors takes an average of $G = 5$ seconds [11].

For the simulation, the travel times along the corridor are a Lognormal distribution with a mean of 0.77 and variance of 0.4 minutes as in Delgado et al. [9], Hernández-Landa et al. [21], but one could use the collected real-date to fine-tune these parameters as in Ricard et al. [28]. For the H-BBP(t_0) model, we consider an average speed of 60 km/h for all buses.

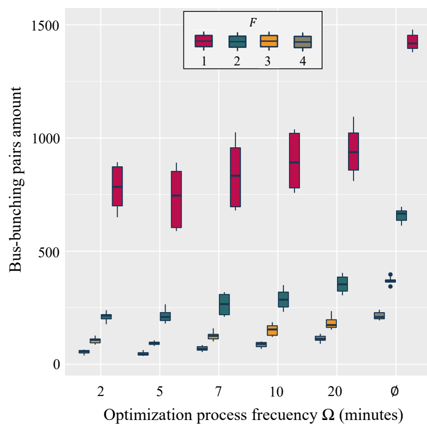
Since we are using a simulation of the Ecovía, we can stress the system to determine high-quality holding policies. Thus, different scenarios are considered to address the following questions (bold parameters are the real case). The following are some critical issues we wish to investigate in our experimental work.

- We wish to investigate if the boarding and alighting times are the leading cause of bus bunching. If this is the case, the companies may evaluate investing in other vehicles with two doors instead of one or implement different queuing processes. Thus, we consider $\underline{U} = \overline{U} \in \{1, \mathbf{2}, 3\}$ seconds per person, as in Tirachini [33].
- We want to know how the bus frequency F impacts the BRT behavior and whether higher frequencies reduce bus bunching. We consider $F \in \{1, \mathbf{2}, 3, 4\}$ buses per minute. $F = 4$ corresponds to the COVID-19 situation where traveling in public transportation was not recommended.
- We wish to establish how many times the holding times must be computed along the planning period. Thus, we test with $\Omega \in \{2, 5, 7, 15, 20, \emptyset\}$, where $\emptyset$ means that no holding times are applied.
- We want to evaluate whether holding times should be an integer number so that drivers adopt them more easily. In addition, we want to know the effect of rounding a holding time, for example, from 35.8 seconds to 30 seconds or one minute. Thus, we consider both cases, the holding times in the H-BBP(t_0) may be integer or continuous values.

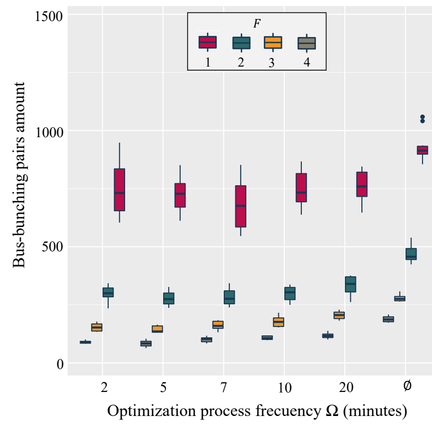
Therefore, $|\underline{U}| \times |F| \times |\Omega| \times 2 = 144$ configurations were tested to evaluate and compare the H-BBP(t_0) model under different operating conditions. Ten simulation runs were carried out for each scenario to guarantee statistical sound results.

4.2 Number of bus bunching pairs

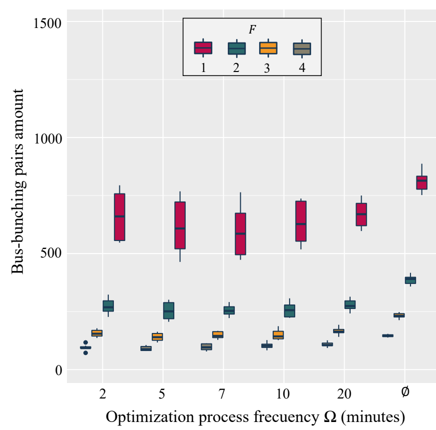
The main objective of the H-BBP(t_0) model is to avoid paired bus bunching. Thus, we apply our BBP methodology to the scenarios with boarding and alighting $\underline{U} = \overline{U} \in \{1, \mathbf{2}, 3\}$ seconds per person corresponding to the three box-plot graphs of Figure 7. The vertical axis indicates the total number of bus bunching pairs throughout the two hours. Note that if there are four buses at a station, then there are three bus bunching. The horizontal axis represents the optimization process's frequency Ω. The four different types of boxes correspond to the different frequencies F.



(a) $\overline{U} = \underline{U} = 1$.



(b) $\overline{U} = \underline{U} = 2$.



(c) $\overline{U} = \underline{U} = 3$.

Figure 7: Bus bunching pairs with respect to optimization process frequency Ω when the bus boarding and alight times are (a) one, (b) two, and (c) three seconds.

Our BBP methodology significantly reduced the number of bus bunching pairs in the evaluation scenarios so users will perceive a more reliable service. Note that generating the holding times more often ($\Omega < 2$) does not necessarily imply more reduction of bus bunching pairs. A period of $\Omega = 5$ or $\Omega = 7$ minutes obtains the best results without overwhelming the drivers with frequent holding requests. On the one hand, when the boarding time of users $\overline{U} = \underline{U}$ is 2 or 3 seconds (Figures 7b and 7c), the least amount of bus bunching pairs occurs when the optimization process Ω is performed every 7 minutes. On the other hand, when the boarding time of the users $\overline{U} = \underline{U}$ is 1 second (Figure 7a), it is recommended that the optimization phase runs every 5 minutes. For systems with a low bus frequency $F = 4$, the bus bunching reduction is not as drastic as for the high-frequency ones of $F = 1$ minute. The case study is observed in Figure 7b with a bus frequency of $F = 2$ without a bus bunching strategy, and there are around 470 bunching pairs. With an intervention every $\Omega = 5$ minutes, we could reduce bus bunching pairs by 45%. The faster boarding and alighting times are, the more bus bunching reductions we get with the BBP methodology. For Figure 7c, corresponding to a COVID-19 situation where line frequency is augmented, we observe that the causes of bus bunching are the boarding and alighting times.

In model H-BBP(t_0), constraints (7) bound the holding times by H minutes. We used $H = 5$ for the previous experimentation of Figure 7, but we carry out other experiments without this bound obtaining similar results about the bus bunching reduction. However, longer holding times could be resented by the users.

4.3 Time-space bus trajectories

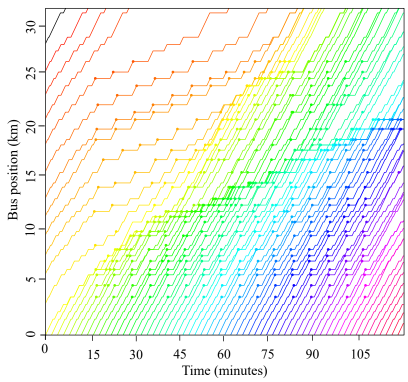
Bus trajectories in time-space graphs allow us to observe the bus bunching cases at a glance whenever two or more lines collide. The four graphs in Figure 8 are similar to the ones in the Section 1. Each line is a bus trip, where the time in minutes is in the horizontal axis while the position of the buses is in the vertical one. Figure 8 corresponds to the case study parameters: boarding and alighting times equal two seconds and bus frequency of two minutes. Graph 8d presents the simulation when no holding times are applied ($\Omega = \emptyset$). We observe many bus bunching pairs and long intervals without buses at the stops, especially after kilometer 10. The optimization stage is done every twenty and ten minutes, respectively, in the time-space graphs 8c and 8b. We can observe a significant change since there are fewer bus bunching cases than without considering holding times. Indeed, the trips tend to be more equidistant, so there are fewer long intervals without buses. We obtain the best solution when $\Omega = 5$ minutes corresponding to Figure 8a. A bus bunching reduction of around 30% is achieved compared to the first scenario.

4.4 User's waiting times

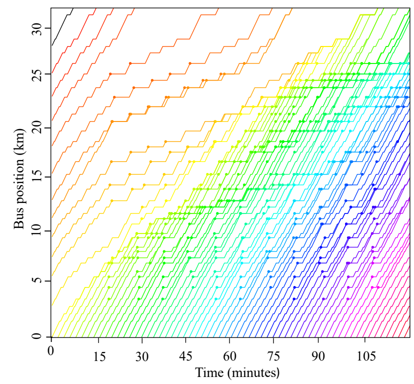
The objective function in the H-BBP(t_0) model seeks to maintain quasi-regular headways to reduce the amount of bus bunching pairs. The main question is if this strategy positively correlates with a reduction of the user's waiting times, which is the essential characteristic of the quality in a public transport system [3, 12].

Figure 9 shows that in our BBP methodology, the minimization of the bus bunching pairs implies a reduction of the user's waiting times. In box-plot graphs of Figure 9, the vertical axis corresponds to the average user's waiting time until they board the first bus. The horizontal axis is the optimization process frequency Ω. As in the previous box-plots of Figure 7, we vary the bus frequency F and the user's boarding and alighting times $\overline{U} = \underline{U}$: (a) one, (b) two, and (c) three seconds.

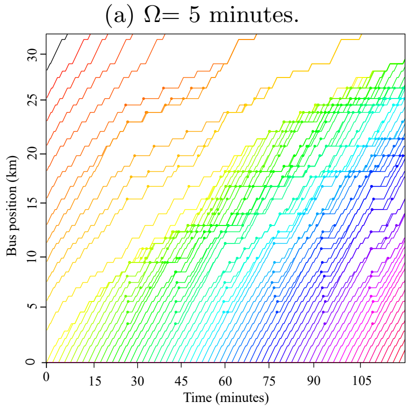
Figures 9a and 9b, show that the more frequent the optimization interventions are, the more the average waiting times are reduced. For the case study (Figure 9b with $F = 2$), the average waiting times are



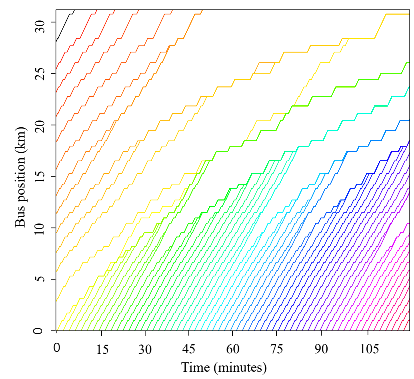
(a) Ω= 5 minutes.



(b) Ω= 10 minutes.



(c) Ω= 20 minutes.

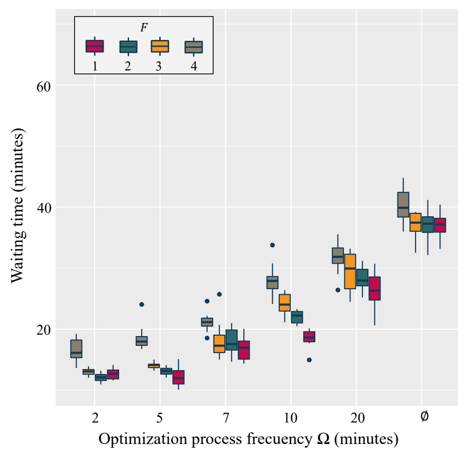


(d) $\Omega = \emptyset$.

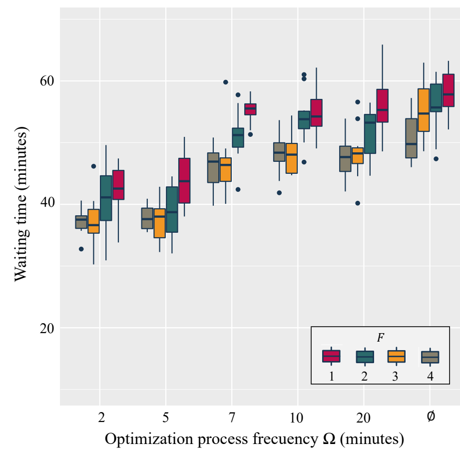
Figure 8: Bus trajectories in two hours with bus boarding and alight times equal to two seconds ($\overline{U} = \underline{U} = 2$), bus frequency of two minutes, and considering different optimization process frequency: (a) every five minutes, (b) every ten minutes, (c) every twenty minutes, and (d) without optimization process.

reduced by 30% for $\Omega = 5$ minutes. The box-plot of Figure 9c shows the case with boarding and alighting times of 3 seconds, which is not usual but observed during the COVID-19 pandemic. In this case, for the bus frequency of 2 and 3 minutes, the waiting times are only reduced when the optimization frequency Ω is small. The variances of the average waiting times are short for the case of 1 second for boarding and alighting (Figure 9a), but they increase as $\overline{U} = \underline{U}$ augment.

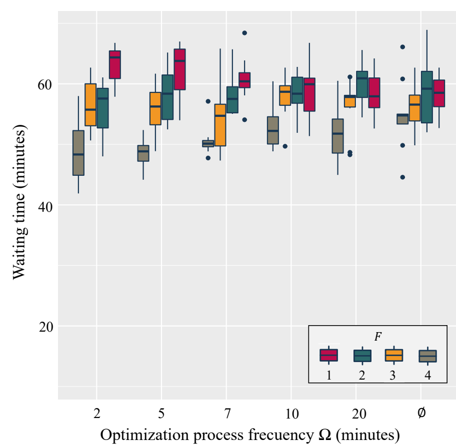
Figure 10 allows us to analyze further the user's average waiting times when the BBP methodology is used. In the figure, there are two percentage ring graphs; both have a bus frequency of $F = 2$, and boarding and alighting waiting times of $\overline{U} = \underline{U} = 2$ corresponding to the case study. The inner ring is without holding times ($\Omega = \emptyset$), while the outer one has a frequency of holding times computation of $\Omega = 7$ minutes. The user's waiting times are classified into five categories: 0-5 minutes, 5-10 minutes, 10-15 minutes, 15-20



(a) $\overline{U} = \underline{U} = 1$.



(b) $\overline{U} = \underline{U} = 2$.



(c) $\overline{U} = \underline{U} = 3$.

Figure 9: Users average waiting times with respect to optimization process frequency Ω when the bus boarding and alighting times are of (a) one, (b) two, and (c) three seconds.

minutes, and more than 20 minutes.

From Figure 10, we observe that computing the holding times via our mathematical model every $\Omega = 7$ minutes positively impacts users' waiting time. Indeed, 15% of the users improved their average waiting times. The bus bunching problem cannot be eradicated with our BBP method, but now only 11% of the passengers wait for more than 20 minutes to board a bus.

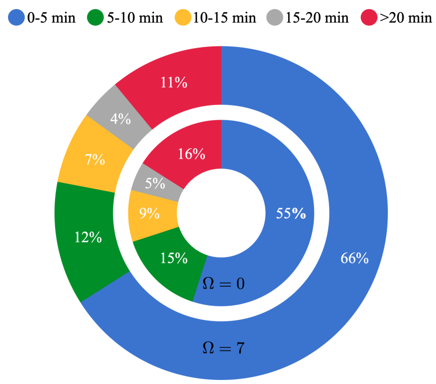


Figure 10: User's average waiting times with respect to optimization process frequency $\Omega = \emptyset$ (inner ring) and $\Omega = 7$ (outer ring) with bus frequency of $F = 2$, and boarding and alighting waiting times of $\overline{U} = \underline{U} = 2$.

4.5 Holding times analysis

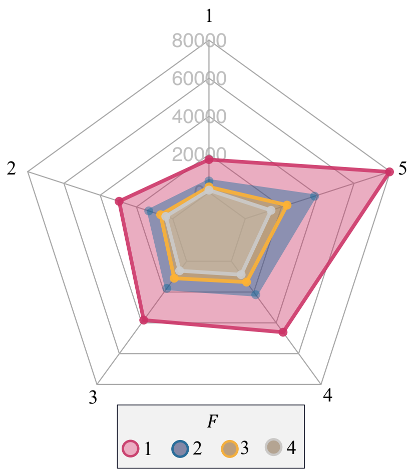
In this section, we analyze if there is a consequence of considering integer holding times which are easier to communicate to the drivers. Thus, Figure 11 considers both cases, the holding times in the H-BBP(t_0) model may be integer (Figure 11b) or continuous values (Figure 11a). In both figures, the pentagon's vertices indicate the different values that the holding times can have. In Figure 11a, vertex $i = 1, \ldots, 4$ groups all values in $[i-1, i)$, while vertex 5 group the values of the holding times between $[4, 5]$. Each internal pentagon represents the number of occurrences of a holding value type differentiated by the different bus frequencies F.

We consider boarding and alighting waiting times of $\overline{U} = \underline{U} = 2$ and all the optimization frequencies of Ω.

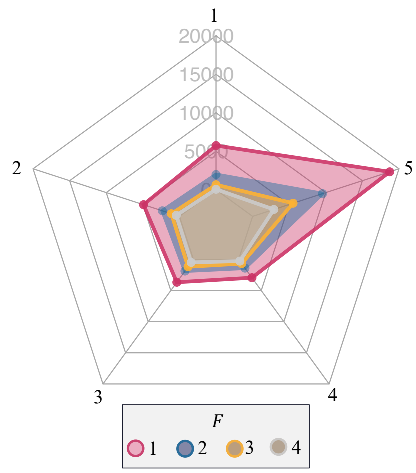
Figure 11 shows that, independently of the nature of the holding times variables in the H-BBP(t_0) model, the value of five minutes is the one with the highest recurrence regardless of service frequency. Figure 11a exhibits holding times have values ranging between [0,60] seconds, which can be challenging to apply in a real case.

We did not notice a significant solving time difference when the holding variables take discrete or continuous values in the H-BBP(t_0) model. On average, the running times for both cases were 15 seconds. Moreover, the number of bus bunching pairs is not affected by the nature of the holding variables. We also analyzed the running times when solving the mathematical model with and without valid inequalities (13) and (14) and considering discrete or continuous holding time variables. A slight improvement in the running time was observed when using valid inequalities, especially with discrete holding variables.

It is also interesting to analyze in which stations the holding times are more recurrent. With this information, we could determine the necessity of applying holding times in all the stations or only at some of them. Figure 12 consists of 12 histograms for which their horizontal axis corresponds to the stops (stop



(a) Continuous holding times variables.



(b) Integer holding times variables.

Figure 11: Values of the holding times variables obtained by H-BBP(t_0) when they are (a) continuous, and (b) discrete.

14 is Mitras) while the vertical one corresponds to the number of times that holding times were needed. We consider the bus frequency F (1,2, 3, and 4 minutes, corresponding to the columns of the figure) and the boarding and alighting times (1, 2, and 3 seconds, corresponding to the lines of the figure). The colors represent the different optimization frequencies Ω. This figure shows the case with integer holding times values.

In Figure 12, the case study is when the bus frequency is $F = 2$ and the boarding and alighting times are two seconds. Before Mitras stop, there is a large number of holding times needed. This behavior holds when $\overline{U} = \underline{U} = 2$ and when $F = 2$ and $F = 3$. Nevertheless, when the boarding and alighting times are short, $\overline{U} = \underline{U} = 1$, the holding times are more uniform through the stops. Thus, the whole Ecovía BRT is affected by the user's boarding and alighting. When the boarding and alighting times are equal to two or three seconds, the tendency is to hold the buses only in the first half of the stops.

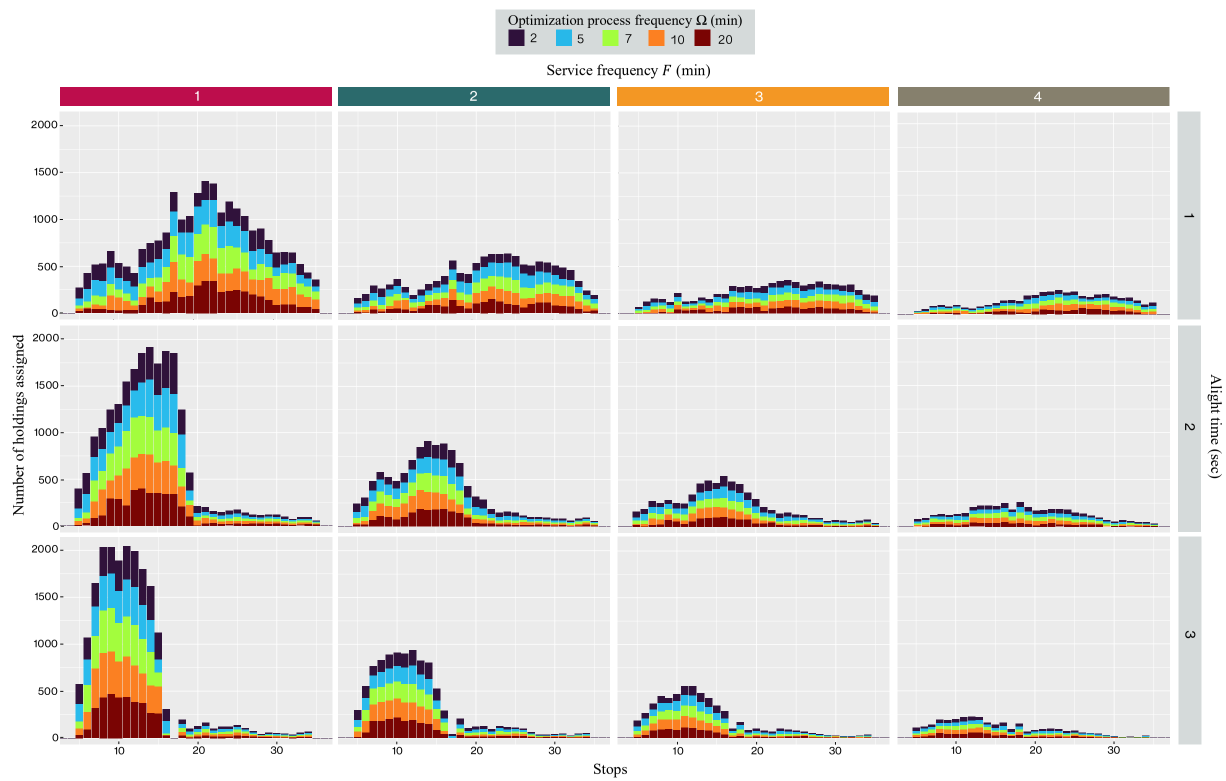


Figure 12: Number of holding times needed at each BRT station.

5 Conclusions

With the BBP methodology, we minimize the bus bunching occurrence; that is, the number of buses of the same line cruising head-to-tail or arriving at a stop simultaneously by determining holding times at the bus stops.

Our BBP solution approach presents an optimization stage that determines the holding times of each bus at the stops by considering the actual information along the bus line and by predicting the future events in the line. This stage is a novel quadratically constrained integer linear programming model that yields the holding times, which are immediately communicated to the drivers to recover quasi-regular headways between the buses. We also present a discrete event simulation program that represents a bus rapid transit system. In this manner, we can retrieve simulated real-time data from the system to validate the efficiency of our mathematical model.

The experimental results show that our methodology mitigates the bus bunching by reducing it up to 45%, especially in the scenarios with high bus frequency. Moreover, the user's waiting times also decrease considerably in scenarios where the dwelling and alighting times are one or two seconds, being reduced by up to 30%. Every five minutes is the frequency in which the holding times must be computed to have the best quasi-equidistant headways between the buses.

Future research lines may consider additional strategies such as bus insertion at stations with a high passenger flow. Also, we could consider a BRT system in which more than one bus line shares stations and some road segments. This framework may be enhanced by considering uncertainty in the user's travel and arrival times and bus speeds, and scenarios where accidents or bus breakdowns may occur. These features will allow us to have more realistic models and more robust solutions. Finally, one may implement heuristic methods to determine holding times for more extensive and more complete BRT systems.

Acknowledgments: The research was supported by the Mexican Council for Science and Technology (grant CONACyT FC 2016-2/1948). The first author was financially supported by graduate scholarship 781117, also by CONACyT.

References

[1] J. Argote-Cabanero, C. F. Daganzo, and J. W. Lynn. Dynamic control of complex transit systems. *Transportation Research Part B: Methodological*, 81:146–160, 2015.

[2] A. Barnett. On controlling randomness in transit operations. *Transportation Science*, 8(2):102–116, 1974.

[3] M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Transportation Studies. MIT Press, Cambridge, 2018.

[4] A. Ceder. *Public Transit Planning and Operation: Modeling, Practice and Behavior*. CRC Press, Boca Raton, 2nd edition, 2016.

[5] C. E. Cortés, D. Sáez, F. Milla, A. Núñez, and M. Riquelme. Hybrid predictive control for real-time optimization of public transport systems' operations based on evolutionary multi-objective optimization. *Transportation Research Part C: Emerging Technologies*, 18(5):757–769, 2010.

[6] C. F. Daganzo. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological*, 43(10):913–921, 2009.

[7] C. F. Daganzo and J. Pilachowski. Reducing bunching with bus-to-bus cooperation. *Transportation Research Part B: Methodological*, 45(1):267–277, 2011.

[8] F. Delgado, J. C. Muñoz, R. Giesen, and A. Cipriano. Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. *Transportation Research Record*, 2090(1):59–67, 2009.

[9] F. Delgado, J. C. Muñoz, and R. Giesen. How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological*, 46(9):1202–1217, 2012.

[10] Y.-J. Deng, X.-H. Liu, X. Hu, and M. Zhang. Reduce bus bunching with a real-time speed control algorithm considering heterogeneous roadway conditions and intersection delays. *Journal of Transportation Engineering, Part A: Systems*, 146(7):04020048, 2020.

[11] K. J. Dueker, T. J. Kimpel, J. G. Strathman, and S. Callas. Determinants of bus dwell time. *Journal of Public Transportation*, 7(1):2–40, 2004.

[12] X. J. Eberlein, N. H. M. Wilson, and D. Bernstein. The holding problem with real–time information available. *Transportation Science*, 35(1):1–18, 2001.

[13] M. Estrada, J. Mensión, J. M. Aymamí, and L. Torres. Bus control strategies in corridors with signalized intersections. *Transportation Research Part C: Emerging Technologies*, 71:500–520, 2016.

[14] R. Fernández, M. A. del Campo, and C. Swett. Data collection and calibration of passenger service time models for the Transantiago system. In *European Transport Conference*, volume 30, Noordwijkerhout, Netherlands, 2008.

[15] L. Fu, Q. Liu, and P. Calamai. Real-time optimization model for dynamic scheduling of transit operations. *Transportation Research Record*, 1857(1):48–55, 2003.

[16] K. Gkiotsalitis and O. Cats. Multi-constrained bus holding control in time windows with branch and bound and alternating minimization. *Transportmetrica B: Transport Dynamics*, 7(1):1258–1285, 2019.

[17] R. Hall, M. Dessouky, and Q. Lu. Optimal holding times at transfer stations. *Computers & Industrial Engineering*, 40(4):379–397, 2001.

[18] S.-X. He. An anti-bunching strategy to improve bus schedule and headway reliability by making use of the available accurate information. *Computers & Industrial Engineering*, 85:17–32, 2015.

[19] S.-X. He, J. Dong, S.-D. Liang, and P.-C. Yuan. An approach to improve the operational stability of a bus line by adjusting bus speeds on the dedicated bus lanes. *Transportation Research Part C: Emerging Technologies*, 107:54–69, 2019.

[20] S.-X. He, S.-D. Liang, J. Dong, D. Zhang, J.-J. He, and P.-C. Yuan. A holding strategy to resist bus bunching with dynamic target headway. *Computers & Industrial Engineering*, 140:106237, 2020.

[21] L. G. Hernández-Landa, M. L. Morales-Marroquín, R. Sánchez-Nigenda, and Y. Á. Ríos-Solís. Linear bus holding model for real-time traffic network control. In *Applied Simulation and Optimization*, pages 303–319. Springer, New York, 2015.

[22] M. D. Hickman. An analytic stochastic model for the transit vehicle holding problem. *Transportation Science*, 35(3):215–237, 2001.

[23] O. J. Ibarra-Rojas, F. Delgado, R. Giesen, and J. C. Muñoz. Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological*, 77:38–75, 2015.

[24] L. M. Leemis and S. K. Park. *Discrete-Event Simulation: A First Course*. Pearson Prentice Hall, Upper Saddle River, 2006.

[25] K. Ling and A. Shalaby. Automated transit headway control via adaptive signal priority. *Journal of Advanced Transportation*, 38(1):45–67, 2004.

[26] H. Liu, A. Skabardonis, and W.-B. Zhang. A dynamic model for tive bus signal priority. In *Proceedings of the 82nd Transportation Research Board Annual Meeting*, Washington, 2003.

[27] E. S. Prassas and R. P. Roess. The highway capacity manual and the committee on highway capacity and quality of service. In *The Highway Capacity Manual: A Conceptual and Research History Volume 2*, volume 12 of *Springer Tracts on Transportation and Traffic*, pages 1–16. Springer, Cham, Switzerland, 2020.

[28] L. Ricard, G. Desaulniers, A. Lodi, and L. M. Rousseau. Predicting the probability distribution of bus travel time to move towards reliable planning of public transport services. Preprint arXiv:2102.02292v1, February 2021.

[29] G. E. Sánchez-Martínez, H. N. Koutsopoulos, and N. H. M. Wilson. Real-time holding control for high-frequency transit with dynamics. *Transportation Research Part B: Methodological*, 83:1–19, 2016.

[30] J. A. Sokolowski and C. M. Banks. *Principles of Modeling and Simulation: A Multidisciplinary Approach*. Wiley, New York, 2011.

[31] A. Sun and M. Hickman. The real–time stop–skipping problem. *Journal of Intelligent Transportation Systems*, 9(2):91–109, 2005.

[32] A. Sun and M. Hickman. The holding problem at multiple holding stations. In *Computer-aided Systems in Public Transport*, volume 600 of *Lecture Notes in Economics and Mathematical Systems*, pages 339–359. Springer, Berlin, Germany, 2008.

[33] A. Tirachini. Bus dwell time: effect of different fare collection systems, bus floor level and age of passengers. *Transportmetrica A: Transport Science*, 9(1):28–49, 2013.

[34] A. Tirachini. The economics and engineering of bus stops: Spacing, design and congestion. *Transportation Research Part A: Policy and Practice*, 59:37–57, 2014.

[35] Transportation Research Board. Transit capacity and quality of service manual. TCRP Report 165, URL: http://www.trb.org/Main/Blurbs/169437.aspx, 2013.

[36] J. Wang and L. Sun. Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework. *Transportation Research Part C: Emerging Technologies*, 116:102661, 2020.

[37] H. P. Williams. *Model Building in Mathematical Programming*. Wiley, New York, 5th edition, 2013.

[38] L. A. Wolsey. *Integer Programming*. Wiley, New York, 2nd edition, 2020.

[39] Y. Xuan, J. Argote, and C. F. Daganzo. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transportation Research Part B: Methodological*, 45(10):1831–1845, 2011.