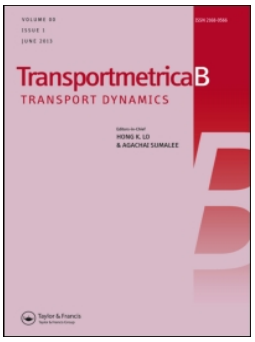

Transportmetrica B: Transport Dynamics

ISSN: 2168-0566 (Print) 2168-0582 (Online) Journal homepage: https://www.tandfonline.com/loi/ttrb20

Multi-constrained bus holding control in time windows with branch and bound and alternating minimization

Konstantinos Gkiotsalitis & Oded Cats

To cite this article: Konstantinos Gkiotsalitis & Oded Cats (2019) Multi-constrained bus holding control in time windows with branch and bound and alternating minimization, Transportmetrica B: Transport Dynamics, 7:1, 1258-1285, DOI: 10.1080/21680566.2019.1606743

To link to this article: https://doi.org/10.1080/21680566.2019.1606743

Published online: 30 Apr 2019.

TRANSPORTMETRICA B: TRANSPORT DYNAMICS
2019, VOL. 7, NO. 1, 1258–1285
https://doi.org/10.1080/21680566.2019.1606743

Taylor & Francis
Taylor & Francis Group

Multi-constrained bus holding control in time windows with branch and bound and alternating minimization

Konstantinos Gkiotsalitis [a*] and Oded Cats [b]

[a]NEC Laboratories Europe, Heidelberg, Germany; [b]Delft University of Technology, GA Delft, The Netherlands

ABSTRACT

This work proposes a periodic bus holding control method where the bus holding times of all running trips are computed simultaneously within each optimization time period; thus, increasing the coordination among running buses for avoiding bus bunching. This paper considers the adverse effects of the bus holding control in the in-vehicle travel times of on-board passengers and performs holistic bus holding decisions by modelling the bus holding problem as a discrete, nonlinear, constrained optimization problem. Given the computational complexity of the bus holding problem, an alternating minimization approach is introduced for computing the optimal holding times at each optimization instance. The performance of the periodic control method is evaluated against the performance of event-based control methods using 5-month automated vehicle location and automated passenger count data from bus line 1 in Stockholm for contacting simulation-based experiments.

ARTICLE HISTORY
Received 22 January 2018
Accepted 9 April 2019

KEYWORDS
Dynamic bus holding; nonlinear programming; periodic control; discrete optimization; bus bunching

1. Introduction

Transport authorities use punctuality-based key performance indicators (KPIs) to evaluate the performance of low-frequency bus services and regularity-based KPIs to evaluate the performance of high-frequency services (Trompet, Liu, and Graham 2011; Gkiotsalitis and Stathopoulos 2016). In the case of high-frequency bus services (i.e. bus lines with more than 5 buses per hour), the most common objective of the bus operations is to maintain an appropriate headway among buses in order to ensure that the waiting times of passengers at stops are close to the planned ones.

In most cases, the dispatching times of bus trips are evenly distributed along the day in order to improve the regularity of services (Ceder 2009; Ceder, Hassold, and Dano 2013). However, given the spatio-temporal variations of traffic and passenger demand, some buses are delayed at some specific segments of the line during the actual operations. These delays have two adverse effects: (i) they increase the headway of the delayed bus from its preceding one; and (ii) they increase the passenger load of the delayed bus that needs to accommodate more passengers who are waiting for an extended time. This causes a 'domino effect' that makes it difficult for a delayed bus to recover and leads to bus bunching, i.e. Gkiotsalitis and Maslekar (2018a) reported that the service regularity deteriorates significantly when the actual travel times differ by more than 30% from their expected values.

Several metropolises with high-frequency bus services such as London, Singapore, Hong Kong, Barcelona (to name a few) have adopted specific KPIs to evaluate the regularity of the running services

CONTACT Konstantinos Gkiotsalitis k.gkiotsalitis@utwente.nl NEC Laboratories Europe, 36 Kurfürsten Anlage, 69115 Heidelberg, Germany; Delft University of Technology, Department of Transport and Planning, P.O. Box 5048, 2600 GA Delft, Netherlands

* Present address: University of Twente Horst-Ring Z-222, P.O. Box 217, 7500 AE Enschede, Netherlands

based on the waiting times of passengers (TfL 2016; LTA 2016). For this purpose, real-time automated vehicle location (AVL) information from the Fleet Telematic System (FTS) of each bus is collected and the recorded arrival times of buses at stops are used for evaluating the performance of the services. To provide an example, the land transport authority (LTA) in Singapore provides bonuses of up to 6000 Singaporean dollars on a monthly basis if the excessive waiting time of passengers is improved by 0.1 min - LTA (2016).

This pressure on bus operators to improve the regularity of their services has led to the introduction of real-time control systems that enable the communication between bus drivers and operational command centers. Nevertheless, most control options are currently based on ad-hoc rules (Cats et al. 2012; Laskaris et al. 2016) which are triggered when a specific event occurs (i.e. a bus arrives late at a stop). Thus, they are not capable of finding a good trade-off between the in-vehicle travel times and the waiting times of passengers at stops as demonstrated in Hans, Chiabaut, and Leclercq (2015). In this study, we investigate this issue and we provide a periodic holding control method that improves the service regularity while considering the in-vehicle travel times.

1.1. Related studies

In general, the daily operations of a bus line are formed on the basis of an extensive tactical planning process that involves the planning of bus frequencies, crew and vehicle schedules (Kepaptsoglou and Karlaftis 2009; Farahani et al. 2013; Gkiotsalitis and Cats 2018; Gkiotsalitis, Wu, and Cats 2019). During the tactical planning phase, the dispatching time of each bus trip is defined and the daily timetable is generated. Nevertheless, tactical planning assumes that the trip travel times remain stable during the actual operations or variate very little from their expected values. This is rarely the case though and several studies have proposed the introduction of real-time corrective actions that can reduce the waiting time variation of passengers at stops. Specifically, these studies have investigated the effect of different corrective measures and can be categorized as:

(1) *Robust slack time planning*. The negative repercussions of corrective actions can cause delays to the dispatching times of future trips that can affect the crew schedules and destabilize the daily plan. Bus operators introduce long slack times to ensure that potential delays are not propagated, thus requiring more buses for maintaining the same frequency level. Daganzo (2009), Zhao, Dessouky, and Bukkapatnam (2006), Xuan, Argote, and Daganzo (2011) and Adamski and Turnau (1998) have studied this problem and have introduced robust optimization methods for calculating efficient slack times which can resist the propagation of delays due to bus holding. These slack times reduce the utilization of buses compared to the case where they are defined based on empirical rules. To compute these optimal slack times, Xuan, Argote, and Daganzo (2011) assumed that the link travel time noise follows the normal distribution and tested the effect of different bus holding options to the overall delay of each trip with the use of simulations.
(2) *Bus Holding control*. The plurality of studies on the bus holding problem (Daganzo 2009; Koehler, Kraus, and Camponogara 2011; Hickman 2001; Eberlein, Wilson, and Bernstein 2001; Laskaris et al. 2018) focus on optimizing the bus holding times at stops for reducing the waiting times of passengers. Such studies have a pre-defined target for the waiting time of passengers at each stop (i.e. scheduled waiting time) and employ heuristics for computing the optimal bus holding times at some bus stops for adhering to the target waiting times. The work of Bartholdi and Eisenstein (2012) is an exception because its objective was to reduce the waiting time variance without adhering to a specific target waiting time. It should be noted here that, as a general practice, buses are not held at every stop because this will increase the passenger inconvenience. In contrary, bus holding is only allowed at a pre-determined sub-set of important bus stops, known as control points.In most of these works, the implications of the bus holding measures to the future trips are not considered and it is assumed that the slack times are long enough to ensure that delays are not propagated. Fu and Yang (2002) tested two of the most common bus

holding strategies: (i) the one-headway-based control where a bus is held at a control point stop based on its time headway with its preceding bus and (ii) the two-headway-based control that considers the headways between its preceding and its following bus. In addition, several works have also considered the travel time noise when computing the bus holding times (Chen, Adida, and Lin 2013; Daganzo 2009). For instance, Hickman (2001) introduced stochasticity at the single holding problem by modeling it as a convex quadratic program in a single variable. Finally, Daganzo and Pilachowski (2011) and Zhao, Bukkapatnam, and Dessouky (2003) proposed distributed control models where buses act as agents that communicate in real-time to achieve dynamic coordination.

(3) *Holding at only one decision point (i.e. last stop)*. Other methods are controlling the buses at one decision point. For instance, in the work of Berrebi, Watkins, and Laval (2015), the buses were not held at control points but only at the last stop (terminal). This problem has been studied since the 1990s (Dessouky et al. 1999; Strathman et al. 1999). In such problems, the holding control is reduced to a dispatching time adjustment problem. This simplification reduces the complexity of the bus holding problem because there is only one decision to be made for each trip.

(4) *Mixed strategies*. Several studies have utilized mixed strategies such as a simultaneous dispatching time and bus holding control (Dessouky et al. 2003; Alesiani and Gkiotsalitis 2018) or introducing stop-skipping, interlining and other measures (Cats et al. 2011, 2012; Diab and El-Geneidy 2012). In particular, the stop-skipping problem has gained significant attention (Furth 1987; Delle Site and Filippi 1998; Fu, Liu, and Calamai 2003; Liu et al. 2013) and mixed strategies that include stop-skipping have been deployed on routes with Origin-Destination (OD) demand peaks at specific segments.

(5) *Holding for Transfer Synchronization*. A distinct line of works has applied bus holding for improving the network synchronization. Hall, Dessouky, and Lu (2001) addressed the real-time bus holding problem for improving synchronization, where bus trips were held at the transfer stops in order to perform the transfer. Hall, Dessouky, and Lu (2001) minimized the transfer times assuming stochastic bus arrivals at transfer stops that follow a normal distribution. Gkiotsalitis and Maslekar (2018b) addressed the holding problem at the dispatching stop for reducing the transfer waiting times and improving the service regularity of each line. In the same direction, Gavriilidou and Cats (2018) devised two control strategies that strive to improve both the single-line regularity and the transfer synchronizations with bus holding. Finally, several other works have addressed the dynamic dispatching time problem considering the objective of increasing the number of synchronizations (Ceder, Golany, and Tal 2001; Ibarra-Rojas and Rios-Solis 2012).

1.2. Focus of this study

Our study is one of the first to examine how a set of holistic holding time decisions can be determined while ensuring that the potential repercussions due to the holding time decisions are taken into consideration. This is a very important issue and was also investigated by Sánchez-Martínez, Koutsopoulos, and Wilson (2016) who produced plans of holding times that accounted not only for the current state of the system, but also for expected travel time and demand changes. For making such holistic holding time decisions, we consider the effects of different holding time combinations and this results in a decision-making process where all holding time options for all running buses are evaluated simultaneously.

This holistic approach increases the complexity of the bus holding control which is no longer event-based (i.e. a holding time is not decided when a bus arrives at one control point stop), but time-window-based (i.e. the holding times of all running buses within a specific time period are decided and updated at the beginning of the time period). For this reason, this work proposes in Sections 2, 3 a space-time model that describes the kinematics of running buses within specific time windows and models the bus holding problem of all running buses as a multi-constrained, non-convex optimization problem. Then, a specific solution method is proposed in Section 4 based on alternating minimization for relaxing the non-convexity of the optimization problem. For providing an exact solution in real-time, a further relaxation is proposed that turns the problem into an Integer Quadratic Programming (IQP) problem and a Branch-and-Bound (B&B) method is introduced.

The main contributions of this work to the state-of-the-art are (i) the introduction of a time-window-based model that makes holistic decisions regarding holding times instead of performing event-based decisions; (ii) the explicit consideration of domino effects due to bus holdings, such as the increase of the in-vehicle travel times and the delay of future trip departures (also known as 'schedule sliding'), with the use of a multi-constrained modeling approach; and (iii) the introduction of a constraint-based model for enabling schedule-based dispatches rather than the headway-based dispatches that are proposed in most works in literature (Daganzo and Pilachowski 2011). The schedule-based dispatches are used because bus operators prefer fixed dispatching times for trips in order to avoid interference with the crew schedules and ensure that many buses, that do not belong exclusively to a particular line, will be available for interlining operations if needed.

To summarize our contribution, this work develops a novel time-window-based model using common assumptions/constraints adopted from past literature and additional objectives and constraints, such as the ones described in (ii)–(iii). Then, we introduce a solution method that can solve the model within a reasonable computational time.

2. Bus kinematic model for a periodic optimization instance

During the daily bus operations, the running buses, $N = \{1, 2, \ldots, n, \ldots, |N|\}$, are expected to serve a number of predefined bus stops. Before proceeding further, it is important to declare the main assumptions of this work:

Assumption 2.1: The number of trips of a bus line are defined during the tactical planning phase and cannot be modified during the actual operations.

Assumption 2.2: Bus operators can apply bus holding control measures at control stops and cannot modify the scheduled dispatching times of buses except in the case of emergency (i.e. due to severe traffic delays).

Assumption 2.3: Overtaking between buses of the same line is not allowed (a reasonable and very common assumption used in most related works Xuan, Argote, and Daganzo 2011; Sun and Hickman 2008; Chen, Adida, and Lin 2013).

Assumption 2.4: Service supply, which is determined at the frequency settings stage, is sufficient for satisfying the passenger demand. This implies that all passengers can be served by the first arriving vehicle even if the number of boardings is increased due to longer headways (see also Eberlein, Wilson, and Bernstein 2001; Hickman 2001).

Assumption 2.5: Our holding method can be applied on high-frequency services (headways of up to 12 min) where we assume that passenger arrivals at stops are random since they cannot coordinate their arrivals with the arrivals of buses (see Osuna and Newell 1972; Hickman 2001; Wu, Liu, and Jin 2016; Vuchic 2017).

Assumption 2.6: Dwell times are negligibly small so that we do not have additional passenger arrivals during the limited time a bus is waiting at a stop (see Marguier 1985; Hickman 2001).

As it will be demonstrated later on, this work applies bus holdings measures that are not expected to postpone the departures of future trips operated by the same buses. This is a significant step towards ensuring that holding will not result in schedule sliding. In addition, this allows the discretization of the daily time horizon into a set of times periods $M = \{\tau, \tau + \delta, \tau + 2\delta, \ldots, \tau + |M|\delta\}$ within which periodic optimizations of the holding measures are allowed to carry out. This rolling horizon optimization approach has been also applied in many studies that focus on stop-skipping (Fu, Liu, and Calamai 2003).

The periodic, time-window-based optimization differs from the event-based headway optimization proposed in most studies (Fu and Yang 2002; Yu and Yang 2009). In a periodic optimization holding control, it is assumed that within a time period δ, the link travel times of trips are available from a link travel time prediction model and the time period δ is short enough so that these travel times are relatively stable. Given the short time window, δ, of rolling horizons, we assume that the stochastic nature of the transit operations can be approximated by a deterministic model because the variance of the stochastic elements becomes very small (Eberlein, Wilson, and Bernstein 2001).

In addition, the periodic optimization control is time-based and an optimization that computes all the holding times of the running buses is applied only at the beginning of the optimization time period $m \in M$. As will be demonstrated in the experimental results in Section 5.1.4, time-window-based optimization is beneficial in scenarios where the travel time prediction errors are up to 50% because, unlike event-based methods, the holistic bus holding decisions can coordination multiple trips.

The expected arrival time of a running bus at one future bus stop, s', during the periodic optimization period $m = \{\tau + m\delta, \tau + (m+1)\delta\}$ is:

$$a_{m,n,s'}(\mathbf{x}) = (\tau + m\delta) + \xi_{\beta_{n,m},\rho_n} + \sum_{i=\rho_n}^{s'-1} t_{m,n,i} + \sum_{i=\rho_n}^{s'-1} d_{m,n,i} + \sum_{i=\rho_n}^{s'-1} x_{m,n,i} \tag{1}$$

where $a_{m,n,s'}$ is the expected arrival time of trip n at stop s' during the mth optimization time period, ρ_n is the next bus stop of trip n given its current position at the beginning of the mth optimization time period and $\xi_{\beta_{n,m},\rho_n}$ is the expected travel time from the position of trip n at the beginning of the optimization, which is denoted as $\beta_{n,m}$, to its next bus stop ρ_n. The expected arrival of the bus at stop s' is affected also by the expected dwell time of trip n at each stop i which is denoted as $d_{m,n,i}$ and the holding times at each stop $x_{m,n,i}$ as it is illustrated in Figure 1.

Figure 1 shows that the expected arrival time of a bus trip at a stop s' is equal to the accumulation of the link travel times, the dwell times and the bus holding times from its current position, $\beta_{n,m}$, until stop s' plus the timestamp $\tau + m\delta$ that denotes the beginning of the optimization time period $m \in M$.

In Equation (1) the link travel times and the dwell times of the bus trip n are modeled as parameters. However, in real-world operations these two parameters have some degree of randomness due to exogenous reasons. Several works such as Shalaby and Farhan (2003), Mazloumi et al. (2011), Chien, Ding, and Wei (2002), Hans et al. (2015) have focused on predicting the travel times of bus operations

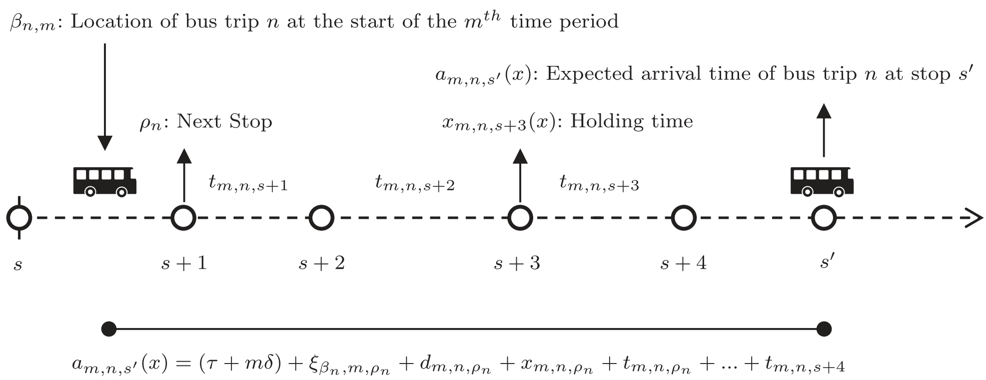

Figure 1. Bus movement illustration.

based on bus data with the use of several techniques spanning from time series analysis to artificial neural networks.

By using the predicted link travel time and dwell time values at the time period mth, the headway between a bus trip n and its preceding trip $n-1$ at stop s can be estimated as:

$$h_{m,n,s}(\mathbf{x}) = a_{m,n,s}(\mathbf{x}) - a_{m,n-1,s}(\mathbf{x}) \tag{2}$$

and its value depends on the applied bus holding measures $\mathbf{x}$. Finally, in high-frequency services where passenger arrivals at stops are random and follow the uniform distribution (Osuna and Newell 1972), the waiting time of passengers who are willing to board on bus trip n is:

$$w_{m,n,s}(\mathbf{x}) = \frac{h_{m,n,s}(\mathbf{x})}{2} = \frac{a_{m,n,s}(\mathbf{x}) - a_{m,n-1,s}(\mathbf{x})}{2} \tag{3}$$

Note that Equation (3) makes use of assumption (2.4) which implies that residual capacity is sufficient so that all waiting passengers are able to board.

3. Model formulation

In this section, we rigorously formulate the mathematical model in order to solve the periodic bus holding optimization problem.

3.1. Parameters and decision variables

Parameters

$S = \{1, 2, \ldots, s, \ldots, \lvert S\rvert\}$	set of bus stops
$N = \{1, \ldots, n, \ldots, \lvert N\rvert\}$	set of daily trips
J	set of bus stops where bus holdings are allowed
υ_n	scheduled dispatching time of each trip $n \in N$
u_n	scheduled arrival time of each trip $n \in N$ at the terminal $\lvert S\rvert$
k_n	embedded slack time for each trip $n \in N$
r_n	required resting time for each trip $n \in N$ before the dispatching of the next trip operated by the same bus
f_n	the next trip after trip n that is operated by the same bus
δ	time duration of each periodic optimization time period
$M = \{\tau, \ldots, \tau + \lvert M\rvert\delta\}$	set of daily periodic optimization time periods
m	identifier of a periodic optimization time period
$\beta_{n,m}$	location of bus trip n at the start of optimization period m
$b_{s,s+1}$	travel distance between bus stops s and $s+1$
$\alpha_{n,s}$	the recorded arrival time of trip n at stop s which is set equal to zero if such information is not available yet
$a_{m,n,s}$	expected arrival time of trip n at stop s during the mth optimization time period without considering bus holdings
$d_{m,n,s}$	expected dwell time of bus trip n at stop s during the mth optimization time period
$t_{m,n,s}$	expected link travel time of trip n between consecutive bus stops s and $s+1$ during the mth optimization time period
ζ_s	desired waiting time of passengers at stop s according to the planned schedule
B_s	passenger arrival rate at stop s

Decision variables

$\mathbf{x} = \{x_{m,n,s}\}$ holding times for each trip $n \in N$ at each stop $s \in S$ during the mth optimization time period

Variables

$a_{m,n,s}(\mathbf{x})$ expected arrival time of trip n at stop s during the mth optimization time period as a function of the bus holding times

$h_{m,n,s}(\mathbf{x})$ headway between bus trips $n-1$ and n at stop s during the mth optimization time period

The aim of this research is to find the optimal holding times, $\mathbf{x}$, for the running buses at each optimization time period. During the mth periodic optimization period, the active and inactive decision variables can be defined by the following inequalities:

$$x_{m,n,s} = \begin{cases} x_{m,n,s} = 0 & \text{if } a_{m,n,s} < \tau + m\delta \vee a_{m,n,s} > \tau + (m+1)\delta \vee s \in J \\ x_{m,n,s} \in q & \text{if } \tau + m\delta \leq a_{m,n,s} \leq \tau + (m+1)\delta \wedge s \notin J \end{cases} \tag{4}$$

where J is the set of control point stops at which a but trip can be held at. Equation (4) contains a series of inequalities that determine whether a bus holding decision variable $x_{m,n,s}$ is equal to zero (inactive) or can take values from a candidate set q. From Equation (4), it is evident that if the expected arrival time of a bus trip n at one stop s is before the start of the mth optimization window $\{\tau + m\delta\}$ or after the end of the mth optimization window $\{\tau + (m+1)\delta\}$, then there is no need to make a decision for that holding time during this optimization time window.

In principle, the admissible set q cannot contain negative values. Many studies such as Lee et al. (2002) have shown that a vehicle driver has a minimum response time of $\simeq 1.5$ s and if a driver is requested to hold a bus at one stop for a very small time (i.e. few milliseconds) he/she will not be able to implement the holding time request with high accuracy. Cortés et al. (2010) has proposed to discretize the holding time instructions to 30-s periods (i.e. 0, 30, 60, ...) seconds. The same 30-s discretization for the holding time decisions is used in commercial software as it is described in Cats et al. (2012). In this study, we propose a higher granularity scheme where the holding time instructions can take values close to the limit of the drivers' responsiveness and, after adding a small time buffer, we propose a 10-s discretization scheme for the holding time options. In such case, the minimum driver response time of $\simeq 1.5$ s can only affect the holding time suggestion by at most 15%. As a result, the set of holding time options is $q = \{0, 10, 20, \ldots\}$ seconds.

3.2. Slack time constraints

During the daily scheduling of the bus operations a slack time is allocated to every trip in order to ensure that the time delay of a bus trip will not postpone the departure of the next trip operated by the same bus. This slack time is a time buffer and is also used for performing bus holdings at control points (Newell 1977; Zhao, Dessouky, and Bukkapatnam 2006).

If a bus trip, n, is expected to arrive at terminal $|S|$ at time $a_{m,n,|S|}(\mathbf{x})$, then the slack time $k_n \in \mathbb{Z}_+$ is the extra time that the bus is allowed to remain at the terminal before starting its next trip.

The expected arrival time of a running trip n at terminal $|S|$, $a_{m,n,|S|}(\mathbf{x})$, can be directly linked to the slack time k_n if we consider the scheduled arrival time of trip n at the terminal, u_n. Satisfying the slack time constraints can be expressed by the following set of inequalities that consider the scheduled arrival time of a trip at the terminal and the associated slack time:

$$a_{m,n,|S|}(\mathbf{x}) \leq u_n + k_n \quad \forall n \in R \tag{5}$$

where R is the set of running trips during the mth optimization period. The inequalities of Equation (5) denote that a bus trip n should be anticipated to arrive at the terminal before its scheduled arrival time, u_n, plus the slack time. We should note here that if due to heavy traffic any of these inequality constraints cannot be satisfied (i.e. the examined trip is delayed beyond the slack time buffer); then, holding this trip at control point stops is no longer allowed.

3.3. Travel time limits for on-board passengers

As explained at the introduction section, this work has a dual objective of maintaining even headways at bus stops in order to reduce bus bunching while at the same time minimizing the negative impact to the on-board passengers. This study imposes strict limits to the travel times of on-board passengers and the number of control point stops that a trip can be held at. Regarding the location of control point stops, several studies such as Abkowitz and Engelstein (1984) proposed that the control point stop should be the stop just preceding a group of stops with high levels of boarding demand. In practice though, the control point stops are generally decided by the transport operators during the tactical planning phase (Shalaby and Farhan 2003; van Oort, Boterman, and van Nes 2012). Therefore, the aim is to ensure that the on-board passenger travel times remain at acceptable levels since the control point stops are, in most cases, predefined.

Let assume that g_n denotes the total bus holding time that is permitted for a trip n. This cumulative limit for the holding times, g_n, can be used for controlling the maximum amount of holding times that are allowed for one trip and can ensure that the on-board passengers will not have a travel time increase of more than g_n due to holdings. Depending on the passenger demand of each trip $n \in N$, the g_n values can vary (i.e. trips with higher demand might be allowed to operate without significant holdings ($g_n \approx 0$), whereas trips with lower demand might be more flexible and allow more time for holdings). The cumulative holding time limits, g_n, per each trip can be provided from the transport authority or derived from historical data analysis regarding the passenger demand levels imposing the following inequality constraints:

$$\sum_{s=1}^{|S|} \sum_{m=1}^{|M|} x_{m,n,s} \leq g_n \quad \forall n \in N \tag{6}$$

One should note that Equation (6) accumulates all the holding times that are calculated during different periodic optimization time periods because the state-space model of this work is implemented in a rolling-horizon.

3.4. Bounding the upper limits of bus holdings

Each holding time should have an upper limit because one bus cannot be held at one stop for a very long time since (i) this increases the on-board passengers' inconvenience; (ii) other buses from the same or different lines should use the same stop for boardings/alightings. Therefore, several studies such as Cortés et al. (2010) and Gkiotsalitis and Maslekar (2015) have proposed to impose an upper limit to the bus holding time which is in the range of 60–90 s depending on the characteristics of the control point stop. Adopting such upper limit for each holding time, the admissible set of holding times is modified to:

$$q = \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90\} \text{ seconds} \tag{7}$$

3.5. Reducing the variance of the actual passenger waiting times from the planned waiting times

Bus operators strive to improve the service reliability by minimizing the passenger waiting time variance from the planned waiting time, ζ_s, at each stop $s \in S$ (Daganzo 2009; LTA 2016). If the scheduled waiting time ζ_s at one stop is different than the actual waiting time of passengers, then the service reliability deteriorates and the bus operators should take corrective actions.

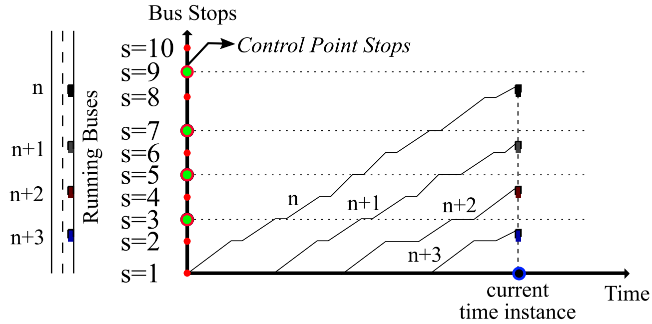

Figure 2. Illustration of the running buses at the beginning of the mth periodic optimization instance.

Let construct an illustrative example where the trips $R = \{n, n+1, n+2, n+3\}$ are the running trips at the start of the mth periodic optimization instance as they are presented in Figure 2. In addition, let $J = \{s = 3, s = 5, s = 7, s = 9\}$ be the set of the control point stops of the bus service in Figure 2. If trips $R = \{n, n+1, n+2, n+3\}$ are the running trips at the mth optimization instance, then trips $R_{\text{prior}} = \{1, 2, \ldots, n-1\}$ have already completed their service prior to the start of the mth optimization time period.

In this way, we can split the daily trips $N = \{1, \ldots, n, \ldots, |N|\}$ into three sets during the mth optimization instance: (i) the trips that are already completed $R_{\text{prior}} = \{1, 2, \ldots, n-1\}$; (ii) the running trips that are expected to operate during the mth time period $R = \{n, n+1, n+2, n+3\}$; and (iii) the remaining trips, $R_{\text{future}} = \{n+4, n+5, \ldots, |N|\}$, until the end of the day.

Having defined the set of running trips, we proceed one step further by identifying the set of trips that are expected to arrive at each bus stop $s \in S$ during the mth optimization time period.

For each bus stop $s \in S$, we initialize a set $R_{m,s}$ that contains all the bus trips that are expected to arrive at that stop during the mth optimization time period. As a result, at the start of the periodic optimization each trip $n \in N$ is assigned to a set $R_{m,s}$ according to the following inequalities:

$$n = \begin{cases} n \in R_{m,s} & \text{if } \alpha_{n,|S|} = 0 \wedge \tau + m\delta \leq a_{m,n,s} \leq \tau + (m+1)\delta \wedge n \in R \\ n \notin R_{m,s} & \text{otherwise} \end{cases} \tag{8}$$

where $\alpha_{n,|S|} = 0$ denotes that trip n has not been completed at the start of the mth optimization time period and $\tau + m\delta \leq a_{m,n,s} \leq \tau + (m+1)\delta$ denotes that the expected arrival time of trip n at stop s lies within the range of the mth optimization time period where:

$$a_{m,n,s} = (\tau + m\delta) + \xi_{\beta_{n,m},\rho_n} + \sum_{i=\rho_n}^{s-1} t_{m,n,i} + \sum_{i=\rho_n}^{s-1} d_{m,n,i} \quad \forall n \in R \tag{9}$$

In Equation (9), $a_{m,n,s}$ is the expected arrival time of trip n at stop s during the mth optimization time period without considering bus holding measures, ρ_n is the next bus stop of trip n given its current position at the beginning of the mth optimization time period and $\xi_{\beta_{n,m},\rho_n}$ is the expected travel time from the position of trip n at the beginning of the optimization, which is denoted as $\beta_{n,m}$, to its next bus stop ρ_n.

After defining the set of trips $R_{m,s}$ that are expected to arrive at stops during the mth optimization time period, the actual waiting time variance from the scheduled waiting times at stops, ζ_s, should be minimized. Using Equation (3), this results in the following objective function:

$$\min_x f(x) = \sum_{s \in S} \sum_{n \in R_{m,s}} \left(\frac{a_{m,n,s}(\mathbf{x}) - a_{m,n-1,s}(\mathbf{x})}{2} - \zeta_s \right)^2 \tag{10}$$

where

$$a_{m,n,s}(\mathbf{x}) = (\tau + m\delta) + \xi_{\beta_{n,m},\rho_n} + \sum_{i=\rho_n}^{s-1} t_{m,n,i} + \sum_{i=\rho_n}^{s-1} d_{m,n,i} + \sum_{i=\rho_n}^{s-1} x_{m,n,i} \quad \forall n \in R. \tag{11}$$

Remark 3.1: As it is evident from Equation (10), the waiting time variance at each stop is equal to the sum of half the headway between consecutive running trips (which is equal to the actual passenger waiting time) subtracted by the scheduled passenger waiting time. One should note though that if n^* is the first element of set $R_{m,s}$, then its preceding trip $n^* - 1$ should be considered at the calculation of the headway even if it does not belong to the set $R_{m,s}$ in order to have a continuity.

The above-mentioned remark ensures that the actual arrival time of the last trip $n^* - 1$ that has arrived at stop s, $(a_{m,n^*-1,s}(\mathbf{x}) = \alpha_{m,n^*-1,s})$, is considered during the periodic optimization and links the past with the current operations (resulting in a no memory-less Markov process).

According to the above descriptions, the first model that tries to (i) minimize the passenger waiting time variation at stops, (ii) satisfy the slack time constraints and (iii) limit the travel times of on-board passengers according to the on-board passenger demand levels can be formulated by combining equations 10, 5, 6 for the optimization of the mth rolling horizon:

$$\begin{aligned}
(Q): \quad & \min_{\mathbf{x}} \quad f(\mathbf{x}) := \sum_{s \in S} \sum_{n \in R_{m,s}} \left(\frac{a_{m,n,s}(\mathbf{x}) - a_{m,n-1,s}(\mathbf{x})}{2} - \zeta_s \right)^2 \\
& \text{subject to} \quad a_{m,n,|S|}(\mathbf{x}) \le u_n + k_n, \quad \forall n \in R \\
& \sum_{s=1}^{|S|} \sum_{m=1}^{|M|} x_{m,n,s} \le g_n, \quad n = 1, \dots, |N|. \\
& x_{m,n,s} = \begin{cases} 0 & \text{if } a_{m,n,s} < \tau + m\delta \text{ or } a_{m,n,s} > \tau + (m+1)\delta \text{ or } s \in J \\ x_{m,n,s} \in q & \text{if } \tau + m\delta \le a_{m,n,s} \le \tau + (m+1)\delta \text{ and } s \notin J \end{cases} \\
& q = \{0, 10, 20, \dots, 90\} \text{ seconds}
\end{aligned} \tag{12}$$

3.6. Modeling the dwell times

The dwell time $d_{m,n,s}$ of a bus trip n at stop s during the mth optimization time period can be associated with the number of boardings and alightings. Many studies, such as Kraft and Bergen (1974), have suggested that the number of boardings has a higher weight on the determination of dwell times. Detailed studies such as Levinson (1983), Guenthner and Sinha (1983) and Bertini and El-Geneidy (2004) showed that there is a linear relationship between the dwell time and the number of boardings where the dwell time can be equal to a minimum value p_0 that can range from 2 to 5 s in case of no boardings and it increases by $p_1 = 1.5 - 4.5$ s for each extra passenger boarding depending on the fare payment structure.

In accordance with the above studies, the underlying relationship between the observed dwell time and the number of boardings can be expressed by a linear equation:

$$d_{m,n,s} = p_0 + p_1 B_{n,s} \tag{13}$$

where p_0 is expressed in seconds, p_1 in seconds per passenger and $B_{n,s}$ is the expected number of passengers that will board on trip n at stop s.

If B_s is the number of passengers per second that are expected to arrive at stop s, the number of passenger boardings, $B_{n,s}$, on one trip n at stop s is associated with the headway between trip n and its preceding trip $n-1$ at that stop. Consequently, the boarding time of passengers $d_{m,n,s}$ is:

$$d_{m,n,s} = p_0 + p_1 B_s(a_{m,n,s} - a_{m,n-1,s}) \tag{14}$$

Equation (14) denotes that the dwell time at one stop is proportional to the headway between two consecutive bus trips $n-1$, n multiplied by the passenger arrival rate at that stop. As a result, holding a running bus $n-1$ at stop s impacts the dwell time of trip n at stop s:

$$d_{m,n,s}(\mathbf{x}) = p_0 + p_1 B_s(a_{m,n,s}(\mathbf{x}) - a_{m,n-1,s}(\mathbf{x})) \tag{15}$$

Plugging Equation (11) and (15) into program (Q) yields:

$$\begin{aligned}
\min_{\mathbf{x}} \quad & f(\mathbf{x}) := \sum_{s\in S}\sum_{n\in R_{m,s}}\Bigg(\frac{1}{2}\Bigg((\tau+m\delta)+\xi_{\beta_{n,m},\rho_n}+\sum_{i=\rho_n}^{s-1} t_{m,n,i}+\sum_{i=\rho_n}^{s-1} d_{m,n,i}(\mathbf{x})+\sum_{i=\rho_n}^{s-1} x_{m,n,i}\Bigg)\\
& -\frac{1}{2}\Bigg((\tau+m\delta)+\xi_{\beta_{n-1,m},\rho_{n-1}}+\sum_{i=\rho_{n-1}}^{s-1} t_{m,n-1,i}+\sum_{i=\rho_{n-1}}^{s-1} d_{m,n-1,i}(\mathbf{x})+\sum_{i=\rho_{n-1}}^{s-1} x_{m,n-1,i}\Bigg)-\zeta_s\Bigg)^2\\
\text{s.t.} \quad & a_{m,n,|S|}(\mathbf{x}) \le u_n + k_n, \quad \forall n \in R\\
& \sum_{s=1}^{|S|}\sum_{m=1}^{|M|} x_{m,n,s} \le g_n, \quad n = 1,\dots,|N|.\\
& x_{m,n,s} = \begin{cases} 0 & \text{if } a_{m,n,s} < \tau+m\delta \text{ or } a_{m,n,s} > \tau+(m+1)\delta \text{ or } s\in J\\ x_{m,n,s}\in q & \text{if } \tau+m\delta \le a_{m,n,s} \le \tau+(m+1)\delta \text{ and } s\notin J \end{cases}\\
& d_{m,n,s}(\mathbf{x}) = p_0 + p_1 B_s(a_{m,n,s}(\mathbf{x}) - a_{m,n-1,s}(\mathbf{x}))\\
& q = \{0,10,20,\dots,90\} \text{ seconds}
\end{aligned} \tag{16}$$

Remark 3.2: From the model of the mth periodic optimization described in Equation (16), one can note that the dwell time of a trip at a stop is affected by the holding times. This results in an open-loop relationship between the arrival times of trips at stops and the dwell times because the dwell time modifications impact the arrival times of trips at stops according to Equation (15) and the arrival time modifications are modifying again the dwell times expressed in Equation (11) leading to a recursive, non-convex optimization problem.

4. Solution method

4.1. Decoupling the arrival and dwell times with alternating minimization

The authors utilize the alternating minimization method for non-convex optimization by solving iteratively the model $(\tilde{Q})$ of Equation (18) where at each iteration the dwell time values are assumed to remain stable resulting in a simpler sub-problem with a convex objective function. The alternating minimization method initializes the problem by assuming that the optimal bus holding times are equal to zero. For this initial solution, $\mathbf{x}^\kappa$, where $x^\kappa_{m,n,s} = 0, \forall x^\kappa_{m,n,s} \in \mathbf{x}^\kappa$, the resulting dwell times are computed:

$$d_{m,n,s} = p_0 + p_1 B_s(a_{m,n,s}(\mathbf{x}^\kappa) - a_{m,n-1,s}(\mathbf{x}^\kappa)) \tag{17}$$

Then, $(\tilde{Q})$ computes the new optimal solution $\mathbf{x}^{\kappa+1}$ under the assumption of fixed dwell time values which were computed by Equation (17). After that, the dwell time values are updated for the bus holding solution $\mathbf{x}^{\kappa+1}$ and the program $(\tilde{Q})$ is solved iteratively until a termination criterion is satisfied.

$$
\begin{aligned}
(\tilde{Q}):\ \min_{\mathbf{x}^{\kappa+1}}\quad & f(\mathbf{x}^{\kappa+1}) := \sum_{s\in S}\sum_{n\in R_{m,s}}\left(\frac{1}{2}\left((\tau+m\delta)+\xi_{\beta_{n,m},\rho_n}+\sum_{i=\rho_n}^{s-1} t_{m,n,i}+\sum_{i=\rho_n}^{s-1} d_{m,n,i}+\sum_{i=\rho_n}^{s-1} x^{\kappa+1}_{m,n,i}\right)\right.\\
& -\frac{1}{2}\left((\tau+m\delta)+\xi_{\beta_{n-1,m},\rho_{n-1}}+\sum_{i=\rho_{n-1}}^{s-1} t_{m,n-1,i}+\sum_{i=\rho_{n-1}}^{s-1} d_{m,n-1,i}\right.\\
& \left.\left.+\sum_{i=\rho_{n-1}}^{s-1} x^{\kappa+1}_{m,n-1,i}\right)-\zeta_s\right)^2\\
\text{s.t.}\quad & a_{m,n,|S|}(\mathbf{x}^{\kappa+1})\le u_n+k_n,\quad \forall n\in R\\
& \sum_{s=1}^{|S|}\sum_{m=1}^{|M|} x^{\kappa+1}_{m,n,s}\le g_n,\quad n=1,\dots,|N|.\\
& x^{\kappa+1}_{m,n,s}=\begin{cases}0 & \text{if } a_{m,n,s}<\tau+m\delta \text{ or } a_{m,n,s}>\tau+(m+1)\delta \text{ or } s\in J\\ x^{\kappa+1}_{m,n,s}\in q & \text{if } \tau+m\delta\le a_{m,n,s}\le\tau+(m+1)\delta \text{ and } s\notin J\end{cases}\\
& d_{m,n,s}=p_0+p_1B_s(a_{m,n,s}(\mathbf{x}^{\kappa})-a_{m,n-1,s}(\mathbf{x}^{\kappa}))\\
& q=\{0,10,20,\dots,90\}\ \text{seconds}
\end{aligned}
\tag{18}
$$

The alternating minimization approach is formed in such a way that, even if the main problem is non-convex, each alternating step in isolation is tractable as summarized in Algorithm 1.

Algorithm 1 Iterative optimization of the holding times at the mth time period

function ALTERNATING MINIMIZATION($\mathbf{x}$)
  Set iteration $\kappa \leftarrow 0$;
  Initialize the problem by setting all holding times to zero for this iteration ($x^{\kappa}_{m,n,s} = 0,\quad \forall x^{\kappa}_{m,n,s}\in\mathbf{x}^{\kappa}$);
  while a convergence criterion is not achieved **do**
    Compute each dwell time $d_{m,n,s}(\mathbf{x}^{\kappa})$ according to Equation 17;
    For each dwell time, set $d_{m,n,s} \leftarrow d_{m,n,s}(\mathbf{x}^{\kappa})$;
    Given the expected dwell times $d_{m,n,s}$ for that holding solution $\mathbf{x}^{\kappa}$, compute a new holding solution $\mathbf{x}^{\kappa+1}$ by solving $(\tilde{Q})$ with Algorithm 2;
    Set $\kappa \leftarrow \kappa+1$ for updating solution $\mathbf{x}^{\kappa}$ with $\mathbf{x}^{\kappa+1}$;
  end while
  return the final solution $\mathbf{x}^{\kappa}$ which is considered as the optimal one
end function

Examining the program $(\tilde{Q})$ that should be solved in every iteration, one can note that the constraints of this optimization problem are linear inequalities and the objective function has a nonlinear form of least absolute deviations.

4.2. Reformulating to an integer quadratic programming (IQP) problem

Note that the bus holding optimization problem of Equation (16) is already reformulated into an alternating minimization problem where at each iteration program $(\tilde{Q})$ should be solved (Algorithm 1).

Plugging Equation (11) into $(\tilde{Q})$ yields:

$$\begin{aligned}
\min_{\mathbf{x}^{\kappa+1}} \quad & f(\mathbf{x}^{\kappa+1}) := \sum_{s\in S}\sum_{n\in R_{m,s}} \left(\frac{1}{2}\left(a_{m,n,s} + \sum_{i=\rho_n}^{s-1} x^{\kappa+1}_{m,n,i}\right) - \frac{1}{2}\left(a_{m,n-1,s} + \sum_{i=\rho_{n-1}}^{s-1} x^{\kappa+1}_{m,n-1,i}\right) - \zeta_s\right)^2 \\
\text{s.t.} \quad & a_{m,n,|S|}(\mathbf{x}^{\kappa+1}) \le u_n + k_n, \quad \forall n \in R \\
& \sum_{s=1}^{|S|}\sum_{m=1}^{|M|} x^{\kappa+1}_{m,n,s} \le g_n, \quad n = 1, \dots, |N|. \\
& x^{\kappa+1}_{m,n,s} = \begin{cases} 0 & \text{if } a_{m,n,s} < \tau + m\delta \text{ or } a_{m,n,s} > \tau + (m+1)\delta \text{ or } s \in J \\ x^{\kappa+1}_{m,n,s} \in q & \text{if } \tau + m\delta \le a_{m,n,s} \le \tau + (m+1)\delta \text{ and } s \notin J \end{cases} \\
& q = \{0, 10, 20, \dots, 90\} \text{ seconds}
\end{aligned} \tag{19}$$

and $(\tilde{Q})$ can be expressed as an Integer Quadratic Programming (IQP) problem:

$$\begin{aligned}
(\tilde{Q}):\ \min_{\mathbf{x}^{\kappa+1}} \quad & f(\mathbf{x}^{\kappa+1}) := \sum_{s\in S}\sum_{n\in R_{m,s}} \left(\frac{1}{4}\left(\sum_{i=\rho_n}^{s-1} x^{\kappa+1}_{m,n,i}\right)^2 + \frac{1}{4}\left(\sum_{i=\rho_n}^{s-1} x^{\kappa+1}_{m,n-1,i}\right)^2\right. \\
& - \frac{1}{2}\left(\sum_{i=\rho_n}^{s-1} x^{\kappa+1}_{m,n,i}\right)\left(\sum_{i=\rho_n}^{s-1} x^{\kappa+1}_{m,n-1,i}\right) \\
& + \left(\frac{a_{m,n,s} - a_{m,n-1,s}}{2} - \zeta_s\right)\sum_{i=\rho_{n-1}}^{s-1} x^{\kappa+1}_{m,n,i} \\
& \left. - \left(\frac{a_{m,n,s} - a_{m,n-1,s}}{2} - \zeta_s\right)\sum_{i=\rho_{n-1}}^{s-1} x^{\kappa+1}_{m,n-1,i} + \left(\frac{a_{m,n,s} - a_{m,n-1,s}}{2} - \zeta_s\right)^2\right) \\
\text{s.t.} \quad & a_{m,n,|S|}(\mathbf{x}^{\kappa+1}) \le u_n + k_n, \quad \forall n \in R \\
& \sum_{s=1}^{|S|}\sum_{m=1}^{|M|} x^{\kappa+1}_{m,n,s} \le g_n, \quad n = 1, \dots, |N|. \\
& x^{\kappa+1}_{m,n,s} = \begin{cases} 0 & \text{if } a_{m,n,s} < \tau + m\delta \text{ or } a_{m,n,s} > \tau + (m+1)\delta \text{ or } s \in J \\ x^{\kappa+1}_{m,n,s} \in q & \text{if } \tau + m\delta \le a_{m,n,s} \le \tau + (m+1)\delta \text{ and } s \notin J \end{cases} \\
& q = \{0, 10, 20, \dots, 90\} \text{ seconds}
\end{aligned} \tag{20}$$

4.3. Solving the discrete bus holding problem with branch and bound

Program $(\tilde{Q})$ expressed in Equation (20) is an NP-Hard, IQP optimization problem. For this reason, we deploy the branch-and-bound (B&B) method that is proposed by Land and Doig (1960) and is one of the most common approaches for solving discrete, NP-Hard problems.

The B&B method explores branches of a dynamically generated tree which represent subsets of the solution set of discrete options. In contrary to the brute-force algorithm that explores the solution space exhaustively, the B&B method uses upper and lower bounds for pruning the set of candidate solutions. In such way, if the estimation of the upper and lower bounds is effective, the exploration effort is significantly reduced.

In order to define the lower bound (LB) of our problem, we allow the holding times to take any real value from 0 to 90 seconds instead of values from the discrete set $q = \{0, 10, 20, \dots, 90\}$. Then, $(\tilde{Q})$

becomes a continuous Quadratic Programming problem (Equation (21)) that can be solved in polynomial time with an exact optimization algorithm such as the Active Set method (for more information regarding the Active Set method, one can refer to Murty and Yu (1988) or Nocedal and Wright (2006)).

$$\begin{aligned} \min_{\mathbf{x}^{\kappa+1}} \quad & f(\mathbf{x}^{\kappa+1}) \\ \text{s.t.} \quad & a_{m,n,|S|}(\mathbf{x}^{\kappa+1}) \leq u_n + k_n, \quad \forall n \in R \\ & \sum_{s=1}^{|S|} \sum_{m=1}^{|M|} x^{\kappa+1}_{m,n,s} \leq g_n, \quad n = 1, \ldots, |N|. \\ & x^{\kappa+1}_{m,n,s} = \begin{cases} 0 & \text{if } a_{m,n,s} < \tau + m\delta \text{ or } a_{m,n,s} > \tau + (m+1)\delta \text{ or } s \in J \\ x^{\kappa+1}_{m,n,s} \in \mathbb{R}_{\geq 0} & \text{if } \tau + m\delta \leq a_{m,n,s} \leq \tau + (m+1)\delta \text{ and } s \notin J \end{cases} \end{aligned} \tag{21}$$

By solving the continuous QP of Equation (21), the lower bound of the problem is established because any discrete solution has inferior performance compared to the global optimum solution of the continuous problem. Initially, the B&B enumeration tree has only one node: the tree root, which is the solution $\mathbf{x}^*$ of the continuous QP problem of Equation (21) and the lower bound is the value of $f(\mathbf{x}^*)$. The objective of the B&B method is to find the optimal discrete solution by efficiently exploring the solution space using upper and lower bounds for pruning the set of candidate solutions.

For exploring the solution space, a typical iteration has three main components: selection of the node to process, branching and bound calculation. If it can be established that the subspace of a node cannot contain the discrete optimal solution, the whole subspace is discarded, else it is stored in the pool of nodes that remain active $A = \Omega - F$ where Ω is the set of all generated nodes and F is the set of discarded nodes.

Initially, the selection of the node to process is trivial because we have only one node (the root $\mathbf{x}^*$). $\mathbf{x}^*$ is the best solution so far, known as *incumbent* solution. However, this is not the solution of our discrete problem $(\tilde{Q})$ of Equation (20) and thus the upper bound of that problem is set to $b^{\text{upper}} \leftarrow +\infty$ because at this stage it is unknown. Then, new branches and nodes are dynamically generated.

During branching, the search space is split into smaller spaces and the problem of Equation (20) is minimized on those smaller spaces for calculating the lower bounds of these spaces (branches). A continuous search space is split into smaller search spaces (branches) by *assigning* a *discrete value* to a variable $x^{\kappa+1}_{m,n',s'} \in \mathbf{x}^{\kappa+1}$. After doing that, the resulting node will have as a lower bound the objective function value of the solution of the restricted continuous QP of Equation (22).

$$\begin{aligned} \min_{\mathbf{x}^{\kappa+1}} \quad & f(\mathbf{x}^{\kappa+1}) \\ \text{s.t.} \quad & a_{m,n,|S|}(\mathbf{x}^{\kappa+1}) \leq u_n + k_n, \quad \forall n \in R \\ & \sum_{s=1}^{|S|} \sum_{m=1}^{|M|} x^{\kappa+1}_{m,n,s} \leq g_n, \quad n = 1, \ldots, |N|. \\ & x^{\kappa+1}_{m,n,s} = \begin{cases} 0 & \text{if } a_{m,n,s} < \tau + m\delta \text{ or } a_{m,n,s} > \tau + (m+1)\delta \text{ or } s \in J \\ x^{\kappa+1}_{m,n,s} \in \mathbb{R}_{\geq 0} & \text{if } \tau + m\delta \leq a_{m,n,s} \leq \tau + (m+1)\delta \text{ and } s \notin J \end{cases} \\ & x^{\kappa+1}_{m,n,s} = x_{m,n',s'} \text{ if } n = n' \text{ and } s = s' \end{aligned} \tag{22}$$

The objective function value of the optimal solution of the restricted continuous QP of a node is its lower bound. That is, if we continue branching from this node the newly generated sub-problems would return inferior objective function values because more of their continuous variables will be discretized.

The node for the new B&B iteration is the node i from the set $A = \Omega - F$ that has the lowest bound value of all other nodes $j \in A$. If after a number of B&B iterations we have a node at which all bus holding times $\mathbf{x}'$ have been assigned values from the discrete set q (that is, $\mathbf{x}' \mid x^{\kappa+1}_{m,n,s} \in q,\ \forall n, s$), then we have a first possible solution for the IQP problem. Note that the performance of this solution, $f(\mathbf{x}')$, is our incumbent upper bound and we set $b^{\text{upper}} \leftarrow f(\mathbf{x}')$.

Then, we can proceed with pruning the solution space. This means that any node that has a higher LB value than the incumbent upper bound, b^{upper}, is excluded from further consideration because branching from this node cannot improve its LB score that is already higher than b^{upper}. In addition, if later on we find another possible IQP solution with lower objective function value than b^{upper}, then this becomes the incumbent discrete solution and the b^{upper} value is updated accordingly. The procedure continues until there are no remaining nodes at the set A which means that all branching possibilities have been exhausted according to the steps of Algorithm 2.

Algorithm 2 Solving the IQP problem with B&B

```
function B&B
    Initialize sets A ← ∅, Ω ← ∅ and F ← ∅;
    Compute the LB of (Q̃) by solving the continuous problem of Equation (21);
    Use this solution as the root of the enumeration tree and add it to sets A, Q;
    while A ≠ ∅ do
        Find the node i ∈ A with the lowest LB score;
        From node i, split the search space into smaller spaces;
        Add the newly generated nodes to the set Ω;
        Compute the LB of each newly generated node by solving the respective QP problem of
        Equation (22);
        Set A ← Ω − F;
        for All discrete solutions in A with variable values from the set q do
            Find the discrete solution with the lowest LB score;
            Set b^upper to the lowest LB score of that discrete solution or to +∞ in case such
            solution does not exist yet;
        end for
        Set F ← F + {i}
        if a node j ∈ A has a LB value higher than b^upper then
            Erase this node by setting F ← F + {j}
        end if
    end while
    return optimal solution of the IQP program (Q̃);
end function
```

5. Numerical experiments

5.1. Small-scale demonstration of one optimization instance

Let consider again the scenario of Figure 2 where at the mth optimization instance we have four running bus trips $n, n{+}1, n{+}2, n{+}3$, whereas $n{-}1$ is the previous trip of trip n that was completed prior to the start of the mth optimization time period. Let also the time duration δ of the periodic optimization to be equal to 10 min (600 s) and the predicted link travel times within this optimization period to be equal to the values of Table 1. Table 1 presents also the expected passenger arrival rate at each stop within the mth optimization time period.

Let the expected travel times of the running trips from their current position to the position of their closest stop to be $\xi_{\beta,n,\rho_n=9} = 35$ s, $\xi_{\beta,n+1,\rho_n=7} = 25$ s, $\xi_{\beta,n+2,\rho_n=5} = 42$ s, $\xi_{\beta,n+3,\rho_n=3} = 47$ s. If the starting time of the mth optimization period is $(\tau + m\delta) \rightarrow 09:10$, or else the 33,000th second from the beginning of the day, it lasts until the 33,600th second of the day. For this 10-minute period,

Table 1. Predicted link travel times, $t_{m,s}$, in seconds and passenger arrival rates, B_s, in passengers per second at each stop s.

	Control point stops, $J = (3,5,7,9)$									
Bus stops, s	1	2	3	4	5	6	7	8	9	10
Arrival rate, B_s	0.4	0.72	0.5	0.76	0.42	0.38	0.37	0.31	0.15	0
Travel times, $t_{m,n+3,s}$	120	130	142	145	138	105	125	105	115	N/A
Travel times, $t_{m,n+2,s}$	120	130	142	145	138	105	125	105	115	N/A
Travel times, $t_{m,n+1,s}$	120	130	142	145	138	105	125	105	115	N/A
Travel times, $t_{m,n,s}$	120	130	142	145	138	105	125	105	115	N/A

Table 2. Arrival times of trips at stops in seconds (the actual arrival times of trips that have already arrived at stops prior to time $\tau + m\delta$ are in bold).

	Bus stops, s									
Trips	1	2	3	4	5	6	7	8	9	10
n+3	**32,688**.5	**32,834**	33,158	33,494.6	*34,014.7*	*34,477.7*	*34,941.8*	*35,485*	*36,014*	*36,373.2*
n+2	**32,338**.5	**32,484**	**32,652**.5	**32,845**.2	**32,998**.7	33,235.5	33,450	*33,686.7*	*33,887.9*	*34,051.5*
n+1	**32,043**	**32,180**	**32,353**.5	**32,540**.7	**32,701**	**32,868**.7	33,065.6	33,294.3	33,500.2	*33,669.6*
n	**31,746**	**31,876**.5	**32,051**	**32,243**	**32,404**	**32,563**.7	**32,710**	**32,882**	33,060.8	33,216.8
n−1	**31,441**	**31,581**	**31,751**	**31,943**	**32,108**	**32,266**	**32,411**	**32,586**	**32,741**	**32,896**

the expected arrival times of buses at stops if we do not consider the impact of the holding times are computed from Equation (11) and presented in Table 2. For the computation of the expected arrival times, the dwell times have also been computed assuming parameter values $p_0 = 5$ s and $p_1 = 1.5$ s/passenger and the passenger arrival rates are presented in Table 1.

In Table 2, the actual arrival times of previous trips are marked with bold to differ from the expected ones during the mth optimization time period. At the same table, all arrival times which are in *italic* denote that the bus trip is not expected to arrive at that stop during the mth optimization time period. For instance, trip n+3 is expected to arrive at stop 4, whereas trip n+2 is expected to arrive at stop 7 and trip n+1 at stop 9. From Table 2, one can infer easily the bus stop arrivals during the mth optimization time period (they are 9 in total). According to Equation (8), the trips that are expected to serve each bus stop within the mth optimization period are: $R_{m,s=1} = \emptyset$, $R_{m,s=2} = \emptyset$, $R_{m,s=3} = \{n+3\}$, $R_{m,s=4} = \{n+3\}$, $R_{m,s=5} = \emptyset$, $R_{m,s=6} = \{n+2\}$, $R_{m,s=7} = \{n+1, n+2\}$, $R_{m,s=8} = \{n+1\}$, $R_{m,s=9} = \{n, n+1\}$, $R_{m,s=10} = \{n\}$.

5.1.1. Waiting time variance without bus holding control

If the scheduled waiting time of passengers at each stop is 2.5 min: $2\zeta_s = 300$ s $\forall s \in S$, then for bus holdings equal to zero (initial solution $\kappa = 1$, where $x^{\kappa}_{m,n,s} = 0, \forall x^{\kappa}_{m,n,s} \in \mathbf{x}^{\kappa}$), the expected waiting time variance of passengers at the mth optimization instance is:

$$f(\mathbf{x}^{\kappa}) = \sum_{s \in S} \sum_{n \in R_{m,s}} \left(\frac{a_{m,n,s}(\mathbf{x}^{\kappa}) - a_{m,n-1,s}(\mathbf{x}^{\kappa})}{2} - \zeta_s \right)^2 = 52954.07\ \mathrm{s}^2$$

5.1.2. Computing the optimal bus holding solution with brute-force

Proceeding with the optimization, the holding times should also ensure that the slack time constraints are satisfied and the on-board passenger travel times will not be significantly affected. Let assume that the scheduled arrival times of bus trips at the last stop u_n, $\forall n \in R$ and the slack times k_n, $\forall n \in R$ are provided by the transport operator and their values are presented in Table 3. Table 3 also presents the upper limits of the accumulated holding times per trip, g_n, given the anticipated on-board passenger levels of each trip.

From Equation (4), the decision variables of the optimization problem are the holding times of the running trips at control point stops $x^{\kappa}_m = (x^{\kappa}_{m,n+3,3}, x^{\kappa}_{m,n+2,7}, x^{\kappa}_{m,n+1,7}, x^{\kappa}_{m,n+1,9}, x^{\kappa}_{m,n,9})$. Starting from the

Table 3. Scheduled arrival times of trips at the last stop, u_n; slack times, k_n; and upper limits for the accumulated holding times for each trip, g_n, in sec.

Trips	u_n in sec.	k_n in sec.	g_n in sec.
$n+3$	36,370	240	300
$n+2$	34,050	240	300
$n+1$	33,680	240	280
n	33,218	240	300

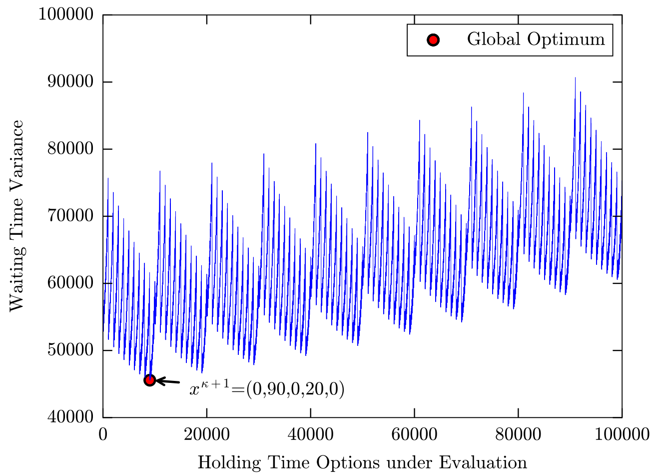

Figure 3. Evaluating $|q|^{|x_m|} = 10^5 = 100{,}000$ bus holding options with Brute Force. The globally optimal solution is (0,90,0,20,0) sec with a performance of 45,571.03 s^2.

initial solution where all holding times were set to zero, $x^{\kappa}_{m,n,s} = 0, \forall x_{m,n,s} \in \mathbf{x}^{\kappa}$, the next solution of the alternating minimization, $\mathbf{x}^{\kappa+1}$, can be computed by solving $(\tilde{Q})$. Before proceeding with such computation, we present in Figure 3 the brute-force solution where the objective function of Equation (16) is evaluated by testing all possible discrete holding time combinations at control point stops from the set $\{0, 10, 20, \ldots, 90\}$ seconds. As it is presented in Figure 3, the number of holding time combinations that are evaluated is exponential, $|q|^{|x_m|}$, where $|x_m|$ is the number of active decision variables at the mth optimization time period. Due to the exponential computational cost, this small-scale scenario requires 100,000 evaluations of the potential bus holding combinations for finding the global optimum; thus, prohibiting the use of brute-force for near real-time optimization.

5.1.3. Demonstrating the computing steps of the alternating minimization method

In this sub-section, we demonstrate the optimization steps of the alternating minimization method that reduces significantly the computational complexity of solving program (Q), enabling its application for periodical bus holding control. The optimization is performed in a computing machine with a 2.40 GHz processor and 16GB RAM and the alternating minimization algorithm presented in Algorithm 1 is programmed in Python 2.7. At each iteration of the alternating minimization, the IQP program $(\tilde{Q})$ is solved with the use of B&B that is also programmed in Python. In addition, the CVXOPT library in Python is used for solving the QP sub-problems according to Algorithm 2.

Initially, the optimal bus holding solution is set equal to $x^{\kappa}_m = (0, 0, 0, 0, 0)$ that results in an objective function score $f(x^{\kappa}) \simeq 52{,}954.07$ as it is presented in Figure 4(a). After that, the dwell times are updated according to Equation (15) and the problem is optimized again with the use of the B&B method of Algorithm 2 resulting in a new solution $x^{\kappa+1}_m = (0, 90, 0, 20, 0)$ with an objective function

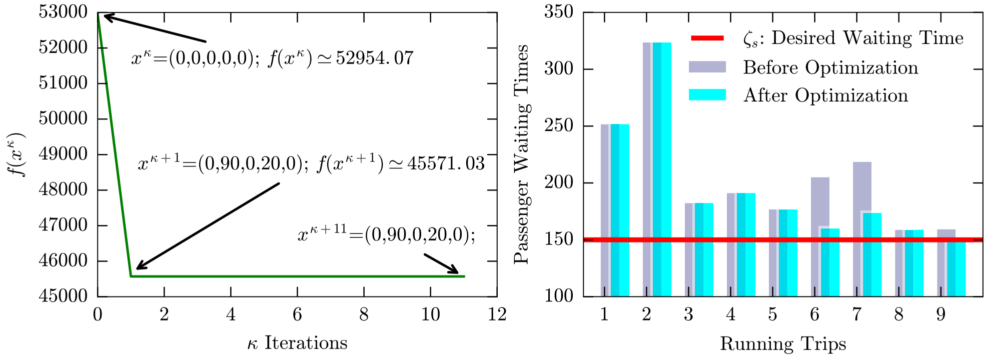

Figure 4. Bus holding control with alternating minimization. (a) Alternating minimization and (b) expected passenger waiting times.

score $f(x^{\kappa+1}) \simeq 45,571.03$. After this step, the dwell times are updated again according to the incumbent solution $x_m^{\kappa+1} = (0, 90, 0, 20, 0)$ and the alternating minimization terminates at the 11th iteration because a further improvement cannot be achieved.

Our alternating minimization approach that solves $(\tilde{Q})$ in each iteration for obtaining a solution of the main program (Q) converges rapidly. In addition, its solution $x_m^{\kappa+11} = (0, 90, 0, 20, 0)$ s which is presented in Figure 4(a) is equivalent to the globally optimal solution of Figure 3 obtained after solving program (Q) with brute-force. Figure 4(b) shows the waiting times of passengers at stops that are expected to be served by the running trips within the mth optimization period: $R_{m,s=1} = \emptyset$, $R_{m,s=2} = \emptyset$, $R_{m,s=3} = \{n+3\}$, $R_{m,s=4} = \{n+3\}$, $R_{m,s=5} = \emptyset$, $R_{m,s=6} = \{n+2\}$, $R_{m,s=7} = \{n+1, n+2\}$, $R_{m,s=8} = \{n+1\}$, $R_{m,s=9} = \{n, n+1\}$, $R_{m,s=10} = \{n\}$. The waiting times of the first five running trips are unaffected by the bus holding control measures because these trips had already served those bus stops before arriving at a control point stop. In contrary, the effect of the bus holding measures is evident at the passenger waiting times for trip n+1 at stop 8 ($R_{m,s=8} = \{n+1\}$), for trips n,n+1 at stop 9 ($R_{m,s=9} = \{n, n+1\}$) and for trip n at stop 10 ($R_{m,s=10} = \{n\}$) resulting in a $\simeq$14% reduction of the passengers' waiting time variance from the desired waiting times.

To demonstrate the optimality gap/computational efficiency of alternating minimization compared to brute-force, we optimize program (Q) in a number of random (idealized) scenarios with different number of running trips. The results are reported in Table 4.

The optimality gap of alternating optimization is negligible (order of magnitude of 0–0.1%) because the alternating minimization relaxes only the dwell times and this relaxation appears to have a small impact on the computation of the key performance indicator. We should note here that dynamic holding decisions have to be made within a matter of several seconds. A brute-force solution will thus become prohibitive when the number of holding decisions exceeds six whereas even ten decisions can be made by the proposed algorithm at the same performance level.

Table 4. Optimality gap and computational cost of alternating optimization compared to brute-force.

	Solution performance (s^2)		Computational cost		
Number of holding decisions	Brute-Force (global optimum)	Alternating optimization	Brute-Force	Alternating optimization	Optimality gap
5	45,571.03	45,571.03	0.53 s	0.34 s	0.00%
6	53,612.06	53,612.06	5.71 s	0.93 s	0.00%
7	58,341.17	58,341.17	50.01 s	1.82 s	0.00%
8	63,572.07	63,572.07	6 min 96 s	2.29 s	0.00%
9	64,671.33	64,675.21	1 h 3 min	3.04 s	0.06%
10	73,458.21	73,479.5	11 h 57 min	4.62 s	0.03%

5.1.4. Comparing the time-window-based optimization against the one-headway-based approach

The proposed periodic bus holding optimization approach can be compared against event-based approaches such as the one-headway-based approach.

Initially, we solve the same problem with the use of the one-headway-based method. By simple inspection, it is evident that when bus trip n arrives at the control point stop $s=9$ it is 33,060.8−32,741 = 319.8 s behind its preceding trip. This results in a passenger waiting time of 159.9 s which is higher than the target waiting time of 150 s. As a result, it is not recommended to hold this trip at stop $s=9$. When bus trip $n+1$ arrives at control point stop $s=9$ it is also 33,500.2−33,060.8 = 439.4 s behind its preceding trip, n, and since 439.4 $> 2\zeta_s$ no holding is applied for this trip also.

Following the same approach for the remaining running trips, the resulting solution of the one-headway-based method is: $\mathbf{x} = (0, 0, 0, 0, 0)$ s. Since all bus trips exhibited higher headways than the target values when they arrived at the control points, no holding measures were applied. This myopic control approach has inferior performance compared to the time-window-based method because there are indeed holding time options that can reduce the variance of passenger waiting times when all running trips are optimized in a holistic manner as demonstrated in Sub-section 5.1.3.

The disadvantage of the time-window-based optimization compared to the one-headway-based approach is that it relies more on the accurate estimation of the link travel times to determine the bus holding times. To conduct a fair comparison, we also consider this aspect by introducing a travel time noise to the near future operations and investigating to what extent the performance of the bus holding times computed by the time-window-based method are affected by this.

For this purpose, the optimal solution of the periodic optimization $\mathbf{x} = (0, 90, 0, 20, 0)$ sec is evaluated with the use of the function:

$$f(\mathbf{x}) = \sum_{s \in S} \sum_{n \in R_{m,s}} \left(\frac{a_{m,n,s}(\mathbf{x}) + \sum_{i=1}^{s-1} \theta_{n,i} - a_{m,n-1,s}(\mathbf{x}) - \sum_{i=1}^{s-1} \theta_{n-1,i}}{2} - \zeta_s \right)^2 \tag{23}$$

where each $\theta_{n,i} \sim N(0, \sigma^2)$ is a link travel time noise for the estimated link travel time $t_{m,n,i}$ that follows the normal distribution with mean value equal to zero and variance $\sigma_{n,i}^2$. Many studies such as Xuan, Argote, and Daganzo (2011) have used the normal distribution to sample the noise of link travel times in order to introduce stochasticity. In this validation setting, we also use the normal distribution and we use the two standard deviations from the mean, $\sigma_{n,i}$, to set the lower and upper bounds of the allowed link travel time variations from the estimated value. Note that two standard deviations from the mean account for 95.45% of the values of a normal distribution.

For our investigation, four main scenarios are generated (one where the standard deviation $\sigma_{n,s}$ of the normal distribution is equal to 10% of the expected link travel time $t_{m,n,s}$ for each trip $n \in N$ and stop $s \in S$, one where it is equal to 20%, one at 50% and one at 80%). At each one of these four scenarios, we run 1,000 simulations where at each simulation the link travel time noise values are sampled from the corresponding normal distribution.

The performance of the periodically optimized bus holding times in the presence of noise is compared against the performance of the one-headway-based solution and the results are summarized in Figure 5. In Figure 5, the Tukey boxplot convention is utilized (plotting the lowest datum still within 1.5 the interquartile range (IQR) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile).

Figure 5 demonstrates the improvement when using the proposed time-window-based holding solution against using the one-headway-based solution in the presence of different link travel time noise levels. The time-window-based optimization approach performs better than the one-headway-based method for travel time noise of 10% and 20%. However, for noise levels of 50% and 80% the improvement is significantly lower demonstrating that when the estimated link travel time values deviate significantly from the actual ones the comparison between the two methods is inconclusive.

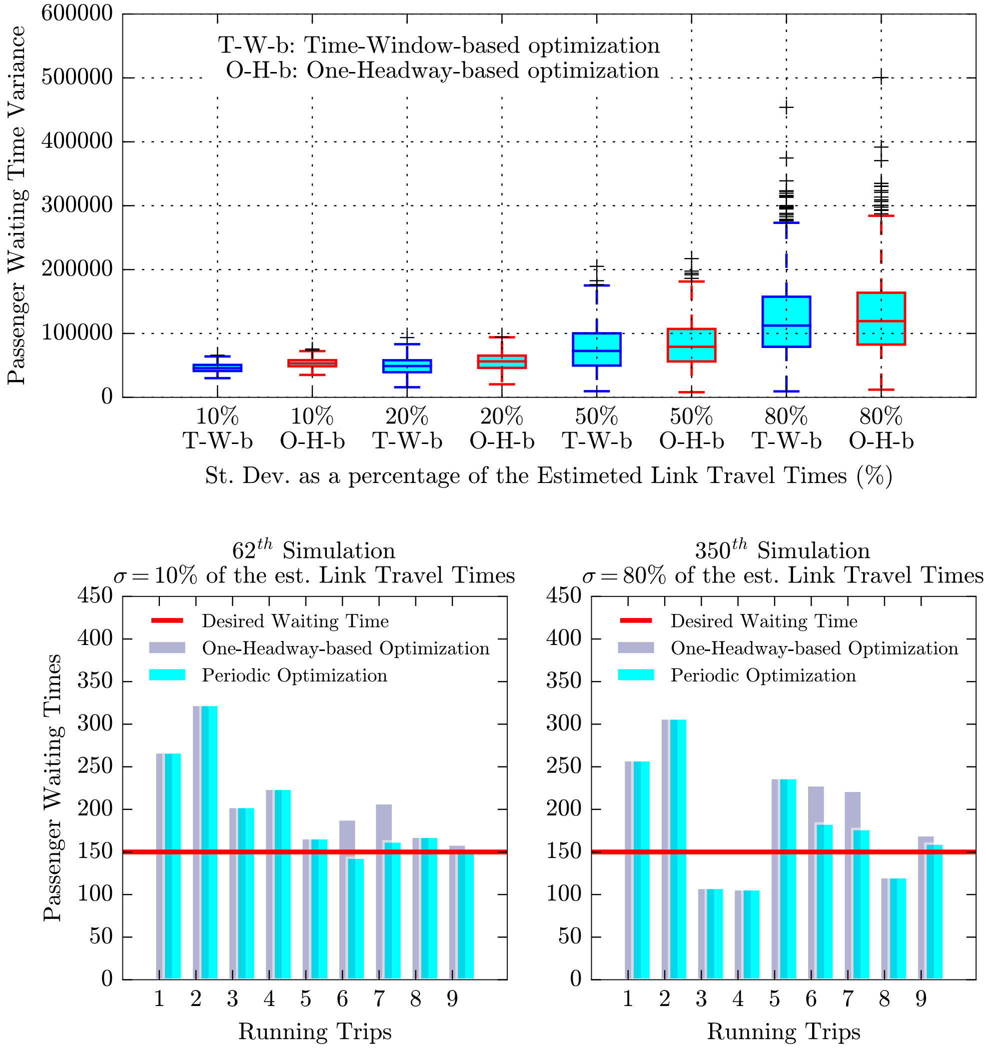

Figure 5. Comparative Analysis of the periodic optimization and the one-headway-based approach in the presence of travel time noise.

Finally, Figure 5 presents also two instances of the results of the one-headway-based bus holding measures and the periodic optimization bus holding measures when the actual link travel times exhibit noise levels of 10% and 80%. An important observation from these two subplots is that the bus holding measures for noise levels of 10% maintain a passenger waiting time variance close to the desired passenger waiting times whereas when the noise levels are in the range of 80% the waiting times vary significantly from their target values due to bus bunching.

5.2. Application: bus line 1 in stockholm

Our application uses data from the eastbound direction of bus line 1 in Stockholm. Line 1 connects the main eastern harbor to a residential area in the western part of the city through the commercial center. Detailed, 5-month datasets of Automatic Vehicle Location (AVL) and Automatic Passenger Counting (APC) data are available for this line that serves 33 bus stops and has four (4) control point stops as presented in Figure 6.

Figure 6. Eastbound direction of Bus Line 1 in Stockholm that serves 33 bus stops from which 4 are control point stops.

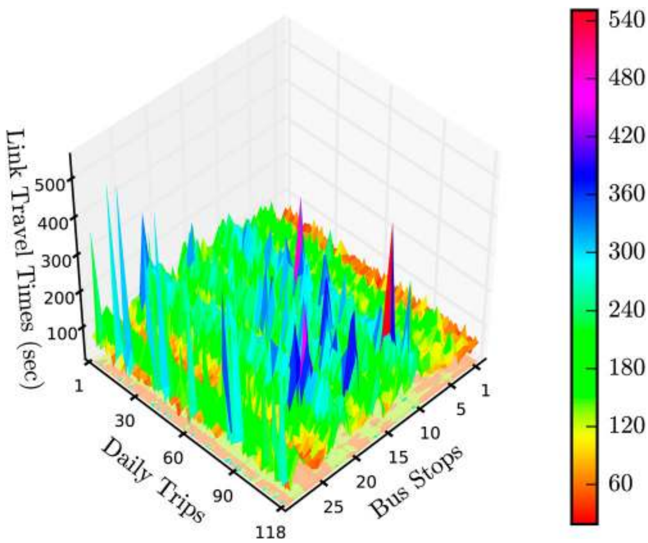

Figure 7. Relation between the estimated link travel times and their predictors (daily trips and bus stops).

The 5-month AVL data is used for estimating the link travel times of the 118 daily trips that operate in the eastbound direction (direction 1) from 07:00 until 19:00. The time period 07:00–19:00 is selected as the main focus area of the daily operations since trips prior to 07:00 and after 19:00 do not experience significant spatio-temporal link travel time variations due to road traffic.

In this work, we utilize the Support Vector Regression (SVR) method for performing a supervised learning task where the target values of the training are the observed link travel times. The SVR method is trained with the use of historical data sets of observed link travel times and returns the expected values $t_{m,n,s}$ presented in Figure 7. If needed, the SVR can be substituted by any other supervised learning method that can be applied to regression problems.

After that, we perform a time-widow-based bus holding control on the daily operations where the entire day is split in 10-min periods within which an optimization occurs. For applying the computed bus holding measures, we import the Stockholm city network in the open source Simulation of Urban MObility (SUMO) using OpenStreetMaps. The AVL data from the actual bus operations of one day was used for replicating the daily operations in SUMO. To this end, we updated automatically the parameters of traffic scenarios with the use of the Constrained Optimization BY Linear Approximation (COBYLA) algorithm from the SciPy library in Python. COBYLA computes the optimal parameter values of traffic flows and traffic signal cycles by minimizing the root-mean-square error (RMSE) between the actual and simulated arrival times of buses at stops.

COBYLA updates the simulation parameter values at each iteration until reaching an acceptable RMSE error that indicates that the simulated arrival times of buses at stops are sufficiently similar to the actual ones, thus allowing us to accept the simulation as a proxy of the real operations. COBYLA was used for calibrating the simulation parameters as instructed by the guidelines of SUMO. Calibrating the simulation parameters is a standard process in SUMO and a benchmark calibration with COBYLA is in Smilowitz et al. (1999). In our case, we adapt this calibration method to our dataset.

The vehicle parameters of the running buses (i.e. acceleration, max. speed, deceleration, vehicle length) are communicated to SUMO using an.xml file that follows the guidelines of SUMO (2017). Table 5 summarizes the main elements of the application that performs $|M| = 72$ periodic optimizations from 07:00 until 19:00 and the vehicle parameter values of the simulated bus trips. In addition, the daily boarding rates expressed as number of boarding passengers per hour are derived from the APC data. For demonstration purposes, the hourly boardings from 08:00 until 12:00 are presented in Figure 8.

The results of the bus holding control in the eastbound direction of line 1 in Stockholm are presented in Figure 9 where the computed bus holding measures at every 10-min optimization horizon are applied in SUMO. In the same figure, the results of the one-headway-based control method are presented where the variance of the passenger waiting times from their target values is evaluated every 10 min.

Table 5. Application characteristics of the bus holding control in bus line 1, direction 1 operating in Stockholm.

Daily trips	Daily period	Time-window	Daily data	Bus holding control	Simulation calibration
$N = 118$	07:00–19:00	$\delta = 10$ min $\|M\| = 72$	AVL & APC	Implemented in SUMO	RMSE minimization of simulated data
				Simulation – Bus parameters: acceleration/deceleration/max. speed	2.61 m/s^2/4.68 m/s^2/100.00 km/h

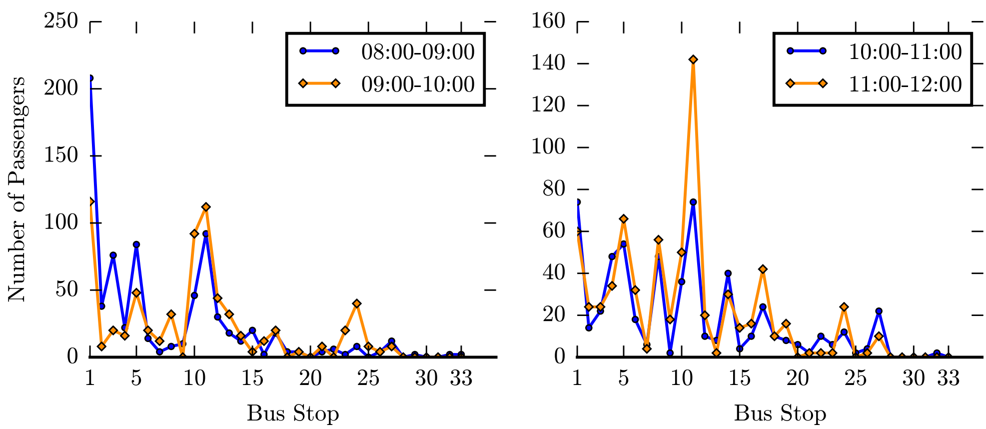

Figure 8. Number of passenger boardings per hour at each bus stop (from 08:00 until 12:00).

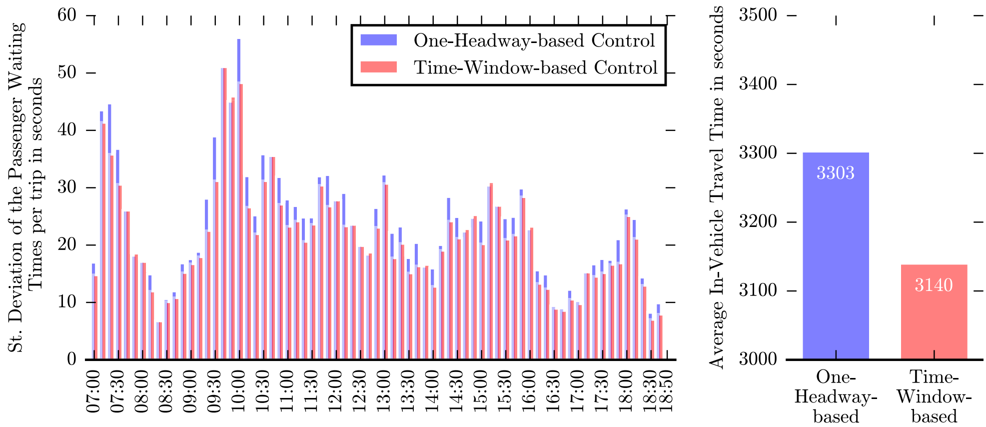

Figure 9. Results of the bus holding control in the eastbound direction of line 1 in Stockholm.

5.3. Discussion

5.3.1. Simulation results

This work evaluated the performance of the bus holding control measures computed by the proposed time-window-based optimization method and the results of each 10-min time window were presented in Figure 9. Furthermore, the one-headway-based bus holding method was also applied in the same scenario resulting in a higher waiting time variance from the desired passenger waiting times. The main reason behind the improved performance of the time-window-based method compared to the event-based methods is that the proposed time-window-based optimization method evaluates the bus holding options holistically within each time window; thus, enabling the coordination of the running buses.

The improvement with the use of the time-window-based optimization method is in the range of 0–17% as it is presented in Figure 9. At some cases, the average waiting time difference of passengers at stops from the desired values was only 3 s whereas in others increased up to 58 s (i.e. during the time period 10:00–10:10).

Even if the evaluation analysis of other works such as Fu and Yang (2002) showed that different event-based methods lead to similar passenger waiting time reductions, this simulation-based analysis demonstrated that the proposed time-window-based method can improve the waiting time variance of passengers by up to $\simeq$17% while improving also the average in-vehicle travel times by $\simeq$5% because the time-window-based optimization process considers the slack time and in-vehicle travel time constraints. Nevertheless, the performance of the time-window-based control might deteriorate when the time window changes (i.e. it is smaller or larger than 10 min). It can also deteriorate when the estimated link travel times differ significantly from their actual values as presented in Figure 5 for noise levels of 50% and more.

5.3.2. Sensitivity analysis of the time interval of a periodic optimization

In the simulation of bus line 1 in Stockholm the time interval of an optimization time window was set to 10 min and the daily period from 07:00 until 19:00 was split into $|M| = 72$ sub-periods. A different time interval has an impact on time-window-based optimization, but does not affect event-based approaches because they only optimize the holding time of one bus at a time. Therefore, we perform a sensitivity analysis of the time-window-based holding solution for different time window intervals and we report the results in Figure 10. This sensitivity analysis is performed by running multiple simulations using the simulation scenario described in Sub-section 5.2 and changing only the time window intervals of the time-window-based control. The external factors of the simulation scenario are the same as in Sub-section 5.2 for limiting the bias in our comparison.

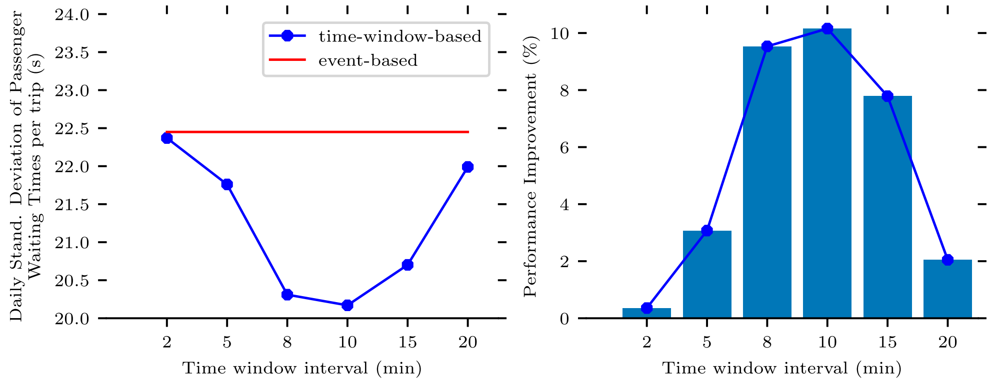

Figure 10. Daily standard deviation of the Passenger Waiting times per trip when applying the time-window-based approach and the one-headway-based approach for different time intervals ranging from 2 to 20 min.

Key observations from Figure 10 are: (i) for a very small time interval of the time-window-based optimization the results are very similar to the event-based approach. This can be explained because in a very small time window (i.e. 2 min) we cannot coordinate too many buses, thus resembling the event-based method which makes holding decisions for one bus at a time; (ii) the waiting time variation is reduced when using time intervals for the periodic optimization in the range of 8–10 min; (iii) increasing the time intervals of the periodic optimization beyond a certain value (10 min in our application) deteriorates the solution performance because our coordinated holdings do not have the opportunity to make use of real-time information to adjust accordingly in case of travel time changes.

5.3.3. Limitations of the proposed time-window-based bus holding method

- Our control method adjusts the holding times of buses at stops and is therefore suitable for correcting the effects of mild disruptions to the service regularity. In the case of severe disruptions, bus operators should consider more radical measures such as changes in the planned service provision and resource allocation (i.e. rescheduling, trip cancellation, short-turning, expressing).
- Our holding method is designed for high-frequency services with headways of up to 12 min.
- Our work is suitable in the context where service supply is sufficient to ensure that there are no passengers who are unable to board due to overcrowding. In contexts where this assumption is invalid, we would have to add the waiting times of stranded passengers to our objective function, resulting in a different optimization problem that will strive to optimize both the headway variability and the bus load.

6. Conclusion

This study proposed a time-window-based bus holding method for improving the waiting time variance of passengers from the planned waiting time values. Time-window-based holding has been proposed by Eberlein, Wilson, and Bernstein (2001) and extended by Delgado, Munoz, and Giesen (2012) and Sánchez-Martínez, Koutsopoulos, and Wilson (2016). Our approach further advances this stream of research by considering the domino effects due to holding with the use of multi-constrained modeling that addresses potential schedule sliding(s). In addition, it introduces a novel solution method using alternating optimization that solves an IQP at each iteration with B&B which proves to return a solution within a reasonable time (Figure 4(a)).

The performance of our approach was evaluated against the one-headway-based method with the use of different travel time noise scenarios. Our holding control exhibited an improved performance with waiting time variance improvements of up to 17% in some specific time periods. Nevertheless, the analysis demonstrated that if the estimated link travel times differ significantly from the actual ones (i.e. noise levels of 50% or more) the computed control measures of the time-window-based approach have an equivalent performance with the control measures of the one-headway-based method.

Although the effect of the travel time noise was considered in the validation phase, the developed time-window-based model does not consider the travel time stochasticity when computing the bus holding measures. Works in event-based bus holding control such as Hickman (2001) have considered the stochasticity of travel times in simple holding models that hold a bus at only one control point stop. However, the complex nature of the time-window-based optimization that requires the simultaneous determination of the holding times of all running trips complicates the computation of holding measures with stochastic optimization and this can be a very interesting topic for future investigation.

Finally, the vast majority of dynamic bus holding works – with the exception of the recent works of Wu, Liu, and Jin (2017, 2018) – have not considered insofar the effect of holding measures on bus loads that might yield increased waiting times to passengers who are unable to board due to overcrowding. Although this effect is not covered in our study and is not considered as a key performance indicator of regularity-based services (Trompet, Liu, and Graham 2011), it is worth to be investigated in future research along the direction of the works of Wu, Liu, and Jin (2017, 2018).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Konstantinos Gkiotsalitis http://orcid.org/0000-0002-3009-1527
Oded Cats http://orcid.org/0000-0002-4506-0459

References

Abkowitz, Mark, and Israel Engelstein. 1984. “Methods for Maintaining Transit Service Regularity.” *Transportation Research Record* 961 (9610): 1–8.

Adamski, Andrzej, and Andrzej Turnau. 1998. “Simulation Support Tool for Real-time Dispatching Control in Public Transport.” *Transportation Research Part A: Policy and Practice* 32 (2): 73–87.

Alesiani, Francesco, and Konstantinos Gkiotsalitis. 2018. “Reinforcement Learning-Based Bus Holding for High-Frequency Services.” In 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 3162–3168. IEEE

Bartholdi, John J., and Donald D. Eisenstein. 2012. “A Self-coördinating Bus Route to Resist Bus Bunching.” *Transportation Research Part B: Methodological* 46 (4): 481–491.

Berrebi, Simon J., Kari E. Watkins, and Jorge A. Laval. 2015. “A Real-time Bus Dispatching Policy to Minimize Passenger Wait on a High Frequency Route.” *Transportation Research Part B: Methodological*81: 377–389.

Bertini, Robert L., and Ahmed M. El-Geneidy. 2004. “Modeling Transit Trip Time using Archived Bus Dispatch System Data.” *Journal of Transportation Engineering* 130 (1): 56–67.

Cats, Oded, Anahid Larijani, Haris Koutsopoulos, and Wilco Burghout. 2011. “Impacts of Holding Control Strategies on Transit Performance: Bus Simulation Model Analysis.” *Transportation Research Record: Journal of the Transportation Research Board* 2216: 51–58.

Cats, O., A. N. Larijani, Á. Ólafsdóttir, W. Burghout, I. Andreasson, and H. N. Koutsopoulos. 2012. “Bus-holding Control Strategies: Simulation-based Evaluation and Guidelines for Implementation.” *Transportation Research Record: Journal of the Transportation Research Board* 2274: 100–108.

Ceder, Avishai. 2009. “Public-transport Automated Timetables using Even Headway and Even Passenger Load Concepts.” In Proceedings of the 32nd Australasian Transport Research Forum (ATRF’09), Auckland, New Zealand, Vol. 17

Ceder, Avishai, B. Golany, and O. Tal. 2001. “Creating Bus Timetables with Maximal Synchronization.” *Transportation Research Part A: Policy and Practice* 35 (10): 913–928.

Ceder, Avishai Avi, Stephan Hassold, and Blanche Dano. 2013. “Approaching Even-load and Even-headway Transit Timetables using Different Bus Sizes.” *Public Transport* 5 (3): 193–217.

Chen, Qin, Elodie Adida, and Jane Lin. 2013. “Implementation of an Iterative Headway-based Bus Holding Strategy with Real-time Information.” *Public Transport* 4 (3): 165–186.

Chien, Steven I-Jy, Yuqing Ding, and Chienhung Wei. 2002. “Dynamic Bus Arrival Time Prediction with Artificial Neural Networks.” *Journal of Transportation Engineering* 128 (5): 429–438.

Cortés, Cristián E., Doris Sáez, Freddy Milla, Alfredo Núñez, and Marcela Riquelme. 2010. "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.

Daganzo, Carlos F.. 2009. "A Headway-based Approach to Eliminate Bus Bunching: Systematic Analysis and Comparisons." *Transportation Research Part B: Methodological* 43 (10): 913–921.

Daganzo, Carlos F., and Josh Pilachowski. 2011. "Reducing Bunching with Bus-to-Bus Cooperation." *Transportation Research Part B: Methodological* 45 (1): 267–277.

Delgado, F., J. C. Munoz, and R. Giesen. 2012. "How Much Can Holding and/or Limiting Boarding Improve Transit Performance?" *Transportation Research Part B: Methodological* 46 (9): 1202–1217.

Delle Site, Paolo, and Francesco Filippi. 1998. "Service Optimization for Bus Corridors with Short-turn Strategies and Variable Vehicle Size." *Transportation Research Part A: Policy and Practice* 32 (1): 19–38.

Dessouky, Maged, Randolph Hall, Ali Nowroozi, and Karen Mourikas. 1999. "Bus Dispatching at Timed Transfer Transit Stations using Bus Tracking Technology." *Transportation Research Part C: Emerging Technologies* 7 (4): 187–208.

Dessouky, Maged, Randolph Hall, Lei Zhang, and Ajay Singh. 2003. "Real-time Control of Buses for Schedule Coordination at a Terminal." *Transportation Research Part A: Policy and Practice* 37 (2): 145–164.

Diab, Ehab I., and Ahmed M. El-Geneidy. 2012. "Understanding the Impacts of a Combination of Service Improvement Strategies on Bus Running Time and Passenger's Perception." *Transportation Research Part A: Policy and Practice* 46 (3): 614–625.

Eberlein, Xu Jun, Nigel H. M. Wilson, and David Bernstein. 2001. "The Holding Problem with Real-Time Information Available." *Transportation Science* 35 (1): 1–18.

Farahani, Reza Zanjirani, Elnaz Miandoabchi, Wai Yuen Szeto, and Hannaneh Rashidi. 2013. "A Review of Urban Transportation Network Design Problems." *European Journal of Operational Research* 229 (2): 281–302.

Fu, Liping, Qing Liu, and Paul Calamai. 2003. "Real-time Optimization Model for Dynamic Scheduling of Transit Operations." *Transportation Research Record: Journal of the Transportation Research Board*1857: 48–55.

Fu, Liping, and Xuhui Yang. 2002. "Design and Implementation of Bus-holding Control Strategies with Real-time Information." *Transportation Research Record: Journal of the Transportation Research Board*1791: 6–12.

Furth, Peter G. 1987. "Short Turning on Transit Routes." *Transportation Research Record* 1108: 42–52.

Gavriilidou, Alexandra, and Oded Cats. 2018. "Reconciling Transfer Synchronization and Service Regularity: Real-time Control Strategies using Passenger Data." *Transportmetrica A: Transport Science*961: 1–29.

Gkiotsalitis, K., and O. Cats. 2018. "Reliable Frequency Determination: Incorporating Information on Service Uncertainty when Setting Dispatching Headways." *Transportation Research Part C: Emerging Technologies* 88: 187–207.

Gkiotsalitis, Konstantinos, and Nitin Maslekar. 2015. "Improving Bus Service Reliability with Stochastic Optimization." In 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), Las Palmas, Spain, 2794–2799. IEEE

Gkiotsalitis, Konstantinos, and Nitin Maslekar. 2018a. "Multiconstrained Timetable Optimization and Performance Evaluation in the Presence of Travel Time Noise." *Journal of Transportation Engineering, Part A: Systems* 144 (9): 04018058.

Gkiotsalitis, Konstantinos, and Nitin Maslekar. 2018b. "Towards Transfer Synchronization of Regularity-based Bus Operations with Sequential Hill-climbing." *Public Transport* 10 (2): 335–361.

Gkiotsalitis, Konstantinos, and Antony Stathopoulos. 2016. "Demand-responsive Public Transportation Re-scheduling for Adjusting to the Joint Leisure Activity Demand." *International Journal of Transportation Science and Technology* 5 (2): 68–82.

Gkiotsalitis, Konstantinos, Zongxiang Wu, and O. Cats. 2019. "A Cost-minimization Model for Bus Fleet Allocation Featuring the Tactical Generation of Short-turning and Interlining Options." *Transportation Research Part C: Emerging Technologies* 98: 14–36.

Guenthner, Richard P., and Kumares C. Sinha. 1983. "Modeling Bus Delays Due to Passenger Boardings and Alightings." *Transportation Research Record* 915: 7–13.

Hall, Randolph, Maged Dessouky, and Quan Lu. 2001. "Optimal Holding Times at Transfer Stations." *Computers & Industrial Engineering* 40 (4): 379–397.

Hans, Etienne, Nicolas Chiabaut, and Ludovic Leclercq. 2015. "Investigating the Irregularity of Bus Routes: Highlighting How Underlying Assumptions of Bus Models Impact the Regularity Results." *Journal of Advanced Transportation* 49 (3): 358–370.

Hans, Etienne, Nicolas Chiabaut, Ludovic Leclercq, and Robert L. Bertini. 2015. "Real-time Bus Route State Forecasting using Particle Filter and Mesoscopic Modeling." *Transportation Research Part C: Emerging Technologies* 61: 121–140.

Hickman, Mark D. 2001. "An Analytic Stochastic Model for the Transit Vehicle Holding Problem." *Transportation Science* 35 (3): 215–237.

Ibarra-Rojas, Omar J., and Yasmin A. Rios-Solis. 2012. "Synchronization of Bus Timetabling." *Transportation Research Part B: Methodological* 46 (5): 599–614.

Kepaptsoglou, Konstantinos, and Matthew Karlaftis. 2009. "Transit Route Network Design Problem: Review." *Journal of Transportation Engineering* 135 (8): 491–505.

Koehler, Luiz Alberto, Werner Kraus, and Eduardo Camponogara. 2011. "Iterative Quadratic Optimization for the Bus Holding Control Problem." *IEEE Transactions on Intelligent Transportation Systems* 12 (4): 1568–1575.

Kraft, Walter H., and Terrence F. Bergen. 1974. “Evaluation of Passenger Service Times for Street Transit Systems.” *Transportation Research Record* 505: 13–20.

Land, Ailsa H., and Alison G. Doig. 1960. “An Automatic Method of Solving Discrete Programming Problems.” *Econometrica* 28: 497–520.

LTA. 2016. “Land Transport Authority: Bus Service Reliability Framework.” Accessed April 21, 2016. http://www.lta.gov.sg/data/apps/news/press/2014/20140124_BSRF(final2)-Annex.pdf.

Laskaris, Georgios, Oded Cats, Erik Jenelius, Marco Rinaldi, and Francesco Viti. 2018. “Multiline Holding Based Control for Lines Merging to a Shared Transit Corridor.” *Transportmetrica B: Transport Dynamics* 1934: 1–34.

Laskaris, Georgios, Oded Cats, Erik Jenelius, and Francesco Viti. 2016. “A Real-time Holding Decision Rule Accounting for Passenger Travel Cost.” In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 2410–2415. IEEE

Lee, John D., Daniel V. McGehee, Timothy L. Brown, and Michelle L. Reyes. 2002. “Collision Warning Timing, Driver Distraction, and Driver Response to Imminent Rear-end Collisions in a High-fidelity Driving Simulator.” *Human Factors* 44 (2): 314–334.

Levinson, Herbert S.. 1983. “Analyzing Transit Travel Time Performance.” *Transportation Research Record* 915: 0–0.

Liu, Zhiyuan, Yadan Yan, Xiaobo Qu, and Yong Zhang. 2013. “Bus Stop-skipping Scheme with Random Travel Time.” *Transportation Research Part C: Emerging Technologies* 35: 46–56.

Marguier, P. H. J. 1985. “Bus Route Performance Evaluation under Stochastic Conditions.” PhD diss., Massachusetts Institute of Technology, Cambridge, MA.

Mazloumi, Ehsan, Geoff Rose, Graham Currie, and Sara Moridpour. 2011. “Prediction Intervals to Account for Uncertainties in Neural Network Predictions: Methodology and Application in Bus Travel Time Prediction.” *Engineering Applications of Artificial Intelligence* 24 (3): 534–542.

Murty, Katta G., and Feng-Tien Yu. 1988. *Linear Complementarity, Linear and Nonlinear Programming*. Vol. 3. Berlin: Heldermann.

Newell, G. F. 1977. “Unstable Brownian Motion of a Bus Trip.” In *Statistical Mechanics and Statistical Methods in Theory and Application*, edited by Uzi Landman, 645–667, Boston, MA: Springer

Nocedal, Jorge, and Stephen J. Wright. 2006. *Sequential Quadratic Programming*. New York, NY, US: Springer.

Osuna, E. E., and G. F. Newell. 1972. “Control Strategies for an Idealized Public Transportation System.” *Transportation Science* 6 (1): 52–72.

Sánchez-Martínez, G. E., H. N. Koutsopoulos, and N. H. M. Wilson. 2016. “Real-time Holding Control for High-frequency Transit with Dynamics.” *Transportation Research Part B: Methodological*83: 1–19.

Shalaby, Amer, and Ali Farhan. 2003. “Bus Travel Time Prediction Model for Dynamic Operations Control and Passenger Information Systems.” *Transportation Research Board 82nd Annual Meeting* 2: 1–16.

Smilowitz, K. R., C. F. Daganzo, M. J. Cassidy, and R. L. Bertini. 1999. “Some Observations of Highway Traffic in Long Queues.” *Transportation Research Record: Journal of the Transportation Research Board* 1678: 225–233.

Strathman, James, Kenneth Dueker, Thomas Kimpel, Rick Gerhart, Ken Turner, Pete Taylor, Steve Callas, David Griffin, and Janet Hopper. 1999. “Automated Bus Dispatching, Operations Control, and Service Reliability: Baseline Analysis.” *Transportation Research Record: Journal of the Transportation Research Board* 1666: 28–36.

SUMO. 2017. “SUMO: Definition of Vehicles, Vehicle Types, and Routes.” Accessed April 12, 2017. http://sumo.dlr.de/wiki/Definition_of_Vehicles,_Vehicle_Types,_and_Routes.

Sun, Aichong, and Mark Hickman. 2008. “The Holding Problem at Multiple Holding Stations.” In *Computer-aided Systems in Public Transport*, edited by Mark Hickman, Pitu Mirchandani and Stefan Voß, 339–359, Springer-Verlag Berlin Heidelberg

TfL. 2016. “Transport for London: Bus Routes and Borough Reports.” Accessed April 21, 2016. https://www.tfl.gov.uk/forms/14144.aspx.

Trompet, Mark, Xiang Liu, and Daniel Graham. 2011. “Development of Key Performance Indicator to Compare Regularity of Service between Urban Bus Operators.” *Transportation Research Record: Journal of the Transportation Research Board* 2216: 33–41.

van Oort, Niels, J. W. Boterman, and Rob van Nes. 2012. “The Impact of Scheduling on Service Reliability: Trip-time Determination and Holding Points in Long-headway Services.” *Public Transport*4 (1): 39–56.

Vuchic, Vukan R. 2017. *Urban Transit: Operations, Planning, and Economics*. Hoboken, New Jersey, US: John Wiley & Sons.

Wu, W., R. Liu, and W. Jin. 2016. “Designing Robust Schedule Coordination Scheme for Transit Networks with Safety Control Margins.” *Transportation Research Part B: Methodological* 93: 495–519.

Wu, Weitiao, Ronghui Liu, and Wenzhou Jin. 2017. “Modelling Bus Bunching and Holding Control with Vehicle Overtaking and Distributed Passenger Boarding Behaviour.” *Transportation Research Part B: Methodological* 104: 175–197.

Wu, Weitiao, Ronghui Liu, and Wenzhou Jin. 2018. “Integrating Bus Holding Control Strategies and Schedule Recovery: Simulation-Based Comparison and Recommendation.” *Journal of Advanced Transportation* 2018: 9407801.

Xuan, Yiguang, Juan Argote, and Carlos F. Daganzo. 2011. “Dynamic Bus Holding Strategies for Schedule Reliability: Optimal Linear Control and Performance Analysis.” *Transportation Research Part B: Methodological* 45 (10): 1831–1845.

Yu, Bin, and Zhongzhen Yang. 2009. "A Dynamic Holding Strategy in Public Transit Systems with Real-time Information." *Applied Intelligence* 31 (1): 69–80.

Zhao, Jiamin, Satish Bukkapatnam, and Maged M. Dessouky. 2003. "Distributed Architecture for Real-time Coordination of Bus Holding in Transit Networks." *IEEE Transactions on Intelligent Transportation Systems* 4 (1): 43–51.

Zhao, Jiamin, Maged Dessouky, and Satish Bukkapatnam. 2006. "Optimal Slack Time for Schedule-based Transit Operations." *Transportation Science* 40 (4): 529–539.