A Bilevel Approach to Frequency Optimization in Public Transportation Systems

Agustín Arizti
Universidad de la República
J. Herrera y Reissig 565, Montevideo, Uruguay
agustin.arizti@fing.edu.uy

Antonio Mauttone
Universidad de la República
J. Herrera y Reissig 565, Montevideo, Uruguay
mauttone@fing.edu.uy

María E. Urquhart
Universidad de la República
J. Herrera y Reissig 565, Montevideo, Uruguay
urquhart@fing.edu.uy

Abstract

We consider the problem of frequency optimization in transit systems, whose objective is to determine the time interval between subsequent buses for a set of public transportation lines. We extend an existing single level model by adding a constraint on bus capacities, while maintaining user choice on routes by means of an assignment sub-model. The resulting formulation is bilevel, and is transformed into a mixed integer linear programming formulation (MILP) that can be solved to optimality for small-sized problem instances, using standard MILP techniques. We study different variants of the same formulation to better understand the bilevel nature of the model and its application to real settings.

2012 ACM Subject Classification Applied computing → Transportation

Keywords and phrases transportation, public transport capacity, transit frequency optimization, mixed integer linear programming, bilevel programming

Digital Object Identifier 10.4230/OASIcs.ATMOS.2018.7

Acknowledgements The authors would like to thank to four reviewers, whose comments have improved the manuscript. This work is partially funded by PEDECIBA.

1 Introduction

There are different stages for the design of a public transportation system based on buses. The literature identifies fives stages [4] that are usually performed sequentially in real systems: route network design, frequency setting, timetable design, fleet assignment and crew assignment. The decisions taken at each stage influence the decisions that can be taken at later stages, and they are taken considering different planning horizons, depending on whether the context of the planning is strategic (long term), tactical (medium term) or operational (short term). The frequency setting decisions are usually part of a tactical planning [10], although at least an initial frequency setting is necessary to evaluate the decisions taken during route network design, which happens on a strategic basis.

The design of a public transportation system needs to consider monetary costs, that range from fixed costs due to the construction of the infrastructure, to variable costs due to the operation of the services. It must also consider the interest of the users, e.g., in providing reasonable travel times, waiting times, and number of transfers. The frequency setting affects directly both concerns, impacting the level of service provided to the users (waiting time, capacity of the lines) and the costs that planners need to incur to run the system (the fleet size is determined by the required frequency of the lines).

The user of a public transportation system usually behaves in an egoistic way, that is, in such a manner as to minimize its individual total travel time (on-board time plus waiting time). Therefore, in order to measure the performance of a transportation system from the viewpoint of the users, models should take into consideration how the users behave when faced with the choice of a specific line from a set of candidate bus lines that can take them to their destinations. Such is the responsibility of an *assignment sub-model*, that by applying a set of hypotheses on how the users behave selects the appropriate lines in order to satisfy travel demands. The assignment model is in itself an optimization problem, usually having a complex formulation and solution method, specially when the influence of the bus capacity is considered in the modelling of the user behavior. Therefore, the complexity of the overall frequency optimization model is strongly determined by the complexity of the underlying assignment sub-model.

In several real settings, public transportation systems run over capacity, meaning that the nominal frequencies of the transit lines are not respected due to lack of capacity. In this context, the capacity is determined by the capacity of the vehicles and the frequencies of the lines. To model these situations, capacity constraints should be taken into consideration when representing the passenger behavior. Even though the problem of transit assignment considering capacities has been properly addressed by the existing literature [15], the problem of transit frequency optimization considering capacities has been more scarcely studied.

The consideration of the bus capacity constraint alongside an assignment sub-model changes the nature of frequency optimization, turning a single level (uncapacitated) formulation into a bilevel one [2]. In bilevel problems there is a constraint that establishes that one or several decision variables must be part of the optimal solution of yet another optimization problem, known as the lower level problem [2] [6]. Exactly two decision makers exist, and the objectives of them do not necessarily coincide. Furthermore, the individual decision each one can take influences the decisions of the other.

The bilevel nature of the frequency optimization problem stems from the fact that the direct addition of bus capacities to the model, involving variables that affect both the planner and the users of the system, would disrupt the underlying assignment sub-model by forcing users to take sub-optimal paths to reach their destination.

The remainder of the article is organized as follows. In section 2 we present a review on related literature and the contributions of this work. The mathematical model and proposed formulation is described in detail in section 3, while in section 4 we present computational experiments using a simple test case on alternative formulations. We conclude the work and refer to future research directions in section 5.

2 Related literature and statement of contribution

In this section we review related relevant literature on frequency optimization in public transportation systems, with a special focus in works that have incorporated either the behavior of the users in an explicit manner (i.e., by means of an assignment sub-model) or bus capacities.

In [8] a nonlinear bilevel formulation for frequency optimization is proposed. It incorporates an explicit assignment model [24] in the lower level, while the upper level problem represents the interest and constraints of the planner, who wants to provide a minimal overall travel time for the users of the system while at the same time diminish the monetary costs by constraining the fleet size. The authors propose a resolution method based on a gradient descent, exploiting specific properties of the problem. The model is applied to several case studies of small to medium sizes.

A bilevel model is proposed in [23], where the upper level seeks to improve an overall cost function and the lower-level consists of the capacity constrained assignment problem formulated in [9]. Tabu Search [17] is used as the heuristic search.

In [20] a MILP formulation is proposed that models user behavior by means of the optimal strategies [24] assignment model. The objective is to minimize the overall travel time of users (on-board travel time plus waiting time) while the operational cost is constrained with an upper limit on the allowed fleet size. The model is solved exactly by using a commercial solver on small instances; for larger instances, a metaheuristic based on Tabu Search is used. The metaheuristic approach is tested using real case studies.

More recently, [18] propose two different integer programming formulations for the problem of designing lines in a public transport system. As part of the line design, frequencies are considered as decision variables to incorporate bus capacities into the model, however, the waiting time of the users is not modeled. Exact solution methods are proposed, and a genetic algorithm is used in order to solve large-scale instances.

The contributions of our work are:

- We consider the passenger behavior as well as the bus capacity and the waiting time of the users, into a single and explicit mathematical programming formulation for the transit frequency optimization problem.
- We propose a bilevel formulation that is converted to a mixed integer linear programming (MILP) formulation suitable of being solved exactly by using commercial MILP solvers for small-sized instances of the problem.
- By applying the exact approach developed to small-sized cases, we are able to study the sensitivity of the solutions with respect to certain aspects of the problem, and therefore, to achieve a better understanding of its nature.

3 Mathematical model

We base our formulation on the one proposed in [20]. In order to model user behavior, it incorporates an explicit assignment model [24].

We propose an extension of the model by adding the bus capacity constraint. This leads us to consider a bilevel formulation that is able to capture the impact that constraints such as the bus capacity, have on the nature of the problem.

3.1 Basic concepts and notation

Before presenting the proposed mathematical programming formulation, we need to provide some concepts as well as a detailed explanation of the used representation.

We make use of a network represented as a directed graph $G = (N, A)$ where nodes acting as bus stops N^P and street endpoints N^S are included in the set N. The movement of the buses along the street is represented by travel arcs (A^T) that connect nodes of N^S. A fixed nonnegative travel time c_a is associated with each travel arc. Boarding (A^B) and alighting (A^L) arcs are also contained in the set A, connecting nodes from N^P to N^S and from N^S to N^P, respectively.

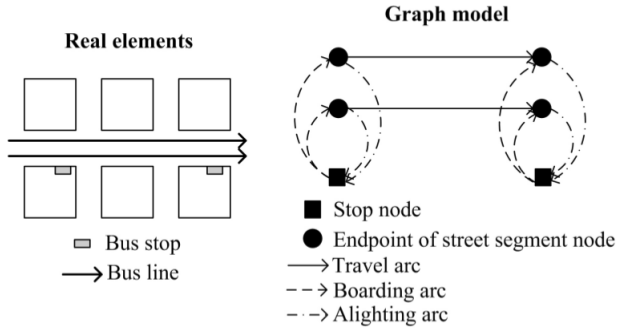

Figure 1 Graph model (extracted from [20]).

We assume that the demand is generated at the bus stops. The demand is represented using an origin-destination matrix, where the set of OD pairs K is such that for a given pair $k \in K$, there are $O_k, D_k \in N^P$ origin and destination nodes, respectively, and a nonnegative value δ_k that represents the amount of people (per time unit in a given time horizon) that have a travel requirement on the pair k.

Lines are defined over the set of travel arcs A^T. Each line $l \in L$ is composed of a sequence of adjacent travel arcs. The round-trip time for a given line is defined as $\sum_{a \in l} c_a$. Lines are either circular, or composed by the concatenation of forward and backward travel arc sequences. Figure 1 illustrates the graph model.

3.2 Assignment model

An assignment model determines user behavior, that is, the way in which users satisfy their travel needs using the existing public transportation lines. Users of the system must choose a line from a set of possible candidate lines that can bring them to their intended destination. Since in order to measure the performance of the system, user satisfaction is of great importance, the assignment model is a critical component of any model of frequency optimization.

The factors that a user considers to make such a choice (i.e., minimize travel time, number of transfers) and the amount of detail and information they have at their disposal (i.e., if the infrastructure provides real time information) determines whether an assignment model is appropriate for the real scenario under study. The way the users behave have a direct influence on the calculation of measures such as the waiting time and occupancy of the buses that end users experience.

The assignment model used in this work is the one proposed in [24], called *optimal strategies*. A strategy is a set of rules that when applied, allow users to reach their destinations. In particular, the model assumes that a given user selects the strategy that minimizes his or her total travel time, including the waiting time at the bus stops. In order to achieve this, it is assumed that users have knowledge of the on-board travel times and frequencies of all the lines of the system. That information is then used to refine a set of attractive lines that can be used to reach the desired destination from the origin. At the bus stop, a given user will take the first bus belonging to the attractive set of lines that passes by that stop. Since the model is probabilistic, an optimal strategy is defined as a strategy that minimizes the total expected travel time.

The probabilistic nature of the model is evident when considering how the waiting time of a passenger waiting on a stop is calculated, for a set of lines $R = \{r_1, \ldots, r_m\}$ with corresponding frequencies $F = \{f_1, \ldots, f_m\}$. As commonly accepted in the literature [10], the

waiting time can then be modeled by a random variable of mean value $E(tw) = \beta / \sum_{r_i \in R} f_i$, where β is a parameter which depends on assumptions concerning service regularity. Since the model assumes that passengers take the first bus that arrives at the stop, the probability of using the route r_i, known as the *frequency share rule*, is $P_i = f_i / \sum_{r_j \in R} f_j$.

For a single OD pair, the assignment model can be formulated as follows:

$$\min_{v,w} \quad \sum_{a \in A} c_a v_a + \sum_{n \in N^P} w_n \tag{1}$$

$$\text{s.t.} \quad \sum_{a \in A_n^+} v_a - \sum_{a \in A_n^-} v_a = b_n \qquad \forall\, n \in N, \tag{2}$$

$$v_a \leq f_a w_n \qquad \forall\, n \in N^P, a \in A_n^+, \tag{3}$$

$$v_a \geq 0 \qquad \forall\, a \in A \tag{4}$$

where w_n is the waiting time multiplied by the amount of demand at node $n \in N^P$, A_n^- are incoming arcs to node n, v_a is the amount of demand flowing through arc $a \in A$, f_a is the frequency of the line corresponding to the boarding arc a, and b_n is a value equal to the demand requirement at that node, that is, δ_k if $n = O_k$, $-\delta_k$ if $n = D_k$, and 0 otherwise.

The objective function (1) states the intention of the users of the system, that is, to minimize their total travel time (sum of on-board travel time and the waiting time at the stops). The flow conservation constraint (2) guarantees that all users are able to reach their destinations. Constraint (3) splits the demand among the different lines that belong to the attractive set, and prohibits flow passing through arc a if the arc is not part of the optimal strategy. If $v_a > 0$ the arc must belong to some optimal strategy and the constraint verifies with equality, restoring the frequency share rule expression.

This is a linear formulation that closely resembles a shortest path problem. The particularities of the formulation consist of a new term in the objective function, representing the waiting time at nodes, and constraint (3) that represents what is known as the *split rule*, where demand is split among the attractive lines leading to the destination and passing by the given stop. Due to the latter constraint, the solution of the assignment problem consists of a *hyperpath* [22] representing different trajectories from origin to destination, instead of a single path on the graph as it is the case when solving the shortest path problem.

The model presented above can be easily extended to consider demand generated (both produced and attracted) in places other than the bus stop. This can be done by considering centroid nodes (representing zones of the study region) which are connected to stop nodes through walking arcs.

3.3 Frequency optimization model

The frequency optimization model proposed in [20] is based on the one proposed in [8], which has a nonlinear bilevel formulation. Formulation (5 - 12) is a linear transformation of that original model, where authors introduce a discretization of the domain of frequencies $\Theta = \{\theta_1 \dots \theta_m\}$ where each element θ_i is a nonnegative value representing a possible value for the frequency of any line.

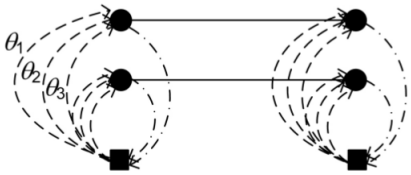

Figure 2 Discretized domain of frequencies (extracted from [20]).

$$\min_{y,v,w} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{5}$$

$$\text{s.t.} \quad \sum_{l\in L}\sum_{f\in\Theta} \theta_f y_{lf} \sum_{a\in l} c_a \leq B, \tag{6}$$

$$\sum_{f\in\Theta} y_{lf} = 1 \qquad \forall\, l \in L, \tag{7}$$

$$\sum_{a\in A_n^+} v_{ak} - \sum_{a\in A_n^-} v_{ak} = b_{nk} \qquad \forall\, n \in N, k \in K, \tag{8}$$

$$v_{ak} \leq \theta_{f(a)} w_{nk} \qquad \forall\, a \in A_n^+, n \in N^P, k \in K, \tag{9}$$

$$v_{ak} \geq 0 \qquad \forall\, a \in A, k \in K, \tag{10}$$

$$v_{ak} \leq \delta_k y_{l(a)f(a)} \qquad \forall\, a \in A^B, k \in K, \tag{11}$$

$$y_{lf} \in \{0,1\} \qquad \forall\, l \in L, f \in \Theta. \tag{12}$$

In doing this, the authors define a new structure of the graph G, where for each line passing by a given bus stop node, there exists as many boarding arcs to that node as possible values of Θ. Figure 2 illustrates the changes introduced in the graph model by using a discretized domain of frequencies.

The model is mixed integer, due to the introduction of the binary variable y_{lf}, which takes value 1 if frequency θ_f is associated with the line l. To keep the planner costs bounded, the parameter B is introduced, which represents an upper limit on the fleet size. To indicate the line frequencies some notation is introduced: $f(a)$ specifies the index in Θ of the frequency associated with the arc a, while $l(a)$ specifies the line that corresponds to that arc. Index k is used to indicate OD pairs.

In formulation (5 - 12) the objective function is that of the users, which intend to minimize their total travel times, while taking into account the interest of the planners that seek to minimize operational costs (6). The assignment model is included in constraints (8 - 10), now expanded to consider each demand pair k. Constraint (7) enforces the fact that each line must have exactly one frequency associated, while constraint (11) prohibits flow on nodes v_{ak} when the frequency associated with that boarding arc is not active ($y_{l(a)f(a)} = 0$) and is redundant otherwise.

This results in a mixed integer linear formulation, where the main source of complexity is the existence of binary variables, and the fact that the discretization of the domain of frequencies increases the size of the underlying graph model due of the addition of new boarding arcs, one per possible frequency value.

3.4 Adding the bus capacity constraint

The assignment sub-model embedded in formulation (5 - 12) assumes that there is sufficient capacity to carry all the passengers that desire to use any line. Furthermore, there is no additional constraint in the formulation that considers the capacity of the lines, which is unrealistic in systems that exhibit high affluence of passengers. Upon introducing a new parameter ω that represents the capacity of a bus, and considering that line capacity (measured in passengers per time unit) is defined as the product of its frequency by the capacity of the bus, we can impose feasible line flows by adding the following constraint:

$$\sum_{k \in K} v_{ak} \leq \sum_{f \in 1..m} y_{l(a)f} \theta_f \omega \quad \forall a \in A^T \tag{13}$$

However, this could result in solutions where the flow of a given OD pair is distributed among:

- A shortest hyperpath comprising lines whose capacity is saturated, i.e., constraint (13) is active for their corresponding travel arcs. This represents the optimal strategy.
- Other alternative hyperpaths, whose cost according to expression (1) is higher than the cost of the shortest one. This represents (sub-optimal) strategies that the users choose a priori, knowing the existence of a shortest hyperpath which is saturated.

This leads us to the concepts of *line planning with route assignment* (LPRA) and *line planning with route choice* (LPRC), first defined in [18]. LPRA models are widespread in the literature, and assume that passengers can be steered by the public transportation planner, an assumption that usually results in simpler but unrealistic models. The utilization of assignment models such as the one used in this work imply a LPRC approach, where each user chooses the route that best fits his or her expectations. Adding constraint (13) directly into the formulation would violate the LPRC approach, as users would need to consider a priori lines that must conform with the new constraint (planners concern) rather than choose the lines in an egoistic way. In a general sense, the addition of any constraint that may impact the variables that model user behavior, and that are not required by the hypothesis of the considered assignment model, would defeat the purpose of the model, since users would behave in a way such as to pursue the optimization of some global optimum that benefits the formulation in place but not necessarily their own interests.

There are at least two ways of modeling the capacity of the buses in the frequency optimization problem while honouring the expected user behavior:

- Assuming that the planner ensures sufficient capacity on the lines that the users want to use. This is done by setting appropriate values of frequencies on the corresponding lines.
- Modeling a congested system, through an assignment sub-model which represents the user behavior under a situation of lack of line capacity. In this case, it is assumed that some users are forced to wait for the next bus of the line, with available capacity, or wait for a different line.

The second one entails to consider an equilibrium assignment sub-model [5] [9] embedded into the frequency optimization model, which is considerably more complex than the first approach [13]. Furthermore, to the best of our knowledge, there is not a formal criterion to decide between both approaches from the modeling point of view. In practice, constraints related to capacity of infrastructure, budget and policy come into play to determine whether it is possible to operate a not congested system. In this work we follow the first approach.

3.4.1 Bilevel mathematical programming formulation

If constraint (13) is added to formulation (5 - 12) we would be considering decisions taken by different actors in the same model. Variables y represent planner decisions in assigning frequencies to lines, while variables v and w represent decisions of the users, that select which lines to use to reach their destinations. Bilevel mathematical programs [2] [6] [12] are used to model scenarios with similar characteristics.

In order to incorporate the bus capacity constraint in our model, we propose the following bilevel formulation:

$$\min_{y,v,w} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{14}$$

$$\text{s.t.} \quad \sum_{l\in L}\sum_{f\in\Theta} \theta_f y_{lf} \sum_{a\in l} c_a \leq B, \tag{15}$$

$$\sum_{f\in\Theta} y_{lf} = 1 \qquad \forall\ l \in L, \tag{16}$$

$$\sum_{k\in K} v_{ak} \leq \sum_{f\in\Theta} y_{l(a)f}\theta_f \omega \qquad \forall\ a \in A^T, \tag{17}$$

$$y_{lf} \in \{0,1\} \qquad \forall\ l \in L, f \in \Theta, \tag{18}$$

$$\min_{v,w} \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{19}$$

$$\text{s.t.} \sum_{a\in A_n^+} v_{ak} - \sum_{a\in A_n^-} v_{ak} = b_{nk} \qquad \forall\ n \in N, k \in K, \tag{20}$$

$$v_{ak} \leq \theta_{f(a)} w_{nk} \qquad \forall\ a \in A_n^+, n \in N^P, k \in K, \tag{21}$$

$$v_{ak} \leq \delta_k y_{l(a)f(a)} \qquad \forall\ a \in A^B, k \in K, \tag{22}$$

$$v_{ak} \geq 0 \qquad \forall\ a \in A, k \in K. \tag{23}$$

where the upper level (14)-(18) represents decisions of the planners while the lower level (19 - 23) represents decisions of the users, that is, the assignment sub-model with the input of fixed frequencies $\theta_{f(a)}$. The objective function of both levels is the same, considering only the objective of the users, which is to minimize the overall travel time. Arguably, the fleet size constraint (15) could be modeled as another objective to minimize at the upper level, which would lead us to consider a multi-objective bilevel formulation, probably increasing the complexity of the formulation [16].

The planners can ensure sufficient capacity on the lines that the users want to use by adjusting the frequencies according to constraint (17). In that manner, users are assumed to perceive unlimited capacities on the lines they might take.

Formulation (14 - 23) is classified as Discrete Continuous Linear Bilevel (DCLB) [2] since the upper level is linear with discrete variables while the lower level is linear with continuous variables. Therefore, it can be reformulated into a MILP problem and in theory it could be solved to optimality. Some commonly used reformulation strategies for doing this are:

- Using the Karush-Kuhn-Tucker (KKT) conditions to substitute the lower level problem and therefore removing the distinction among the different levels. Due to the complementarity term, that is not linear, the resulting reformulation would be a standard single level nonlinear mathematical program that is suitable to be solved by some of the existing nonlinear algorithms. Usually, the reformulation is combined with a linearization of the complementary slackness term using the *big-M* method [12]. This approach has been described and used in [2] [12].

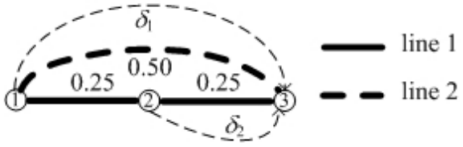

Figure 3 Illustrative example.

- Primal-Dual reformulation. In this case the lower level problem is replaced by using its dual constraints, primal (original) constraints, and the strong duality theorem equality (equality between the lower and upper level objective functions), since the KKT conditions are equivalent to the later conditions when the lower level problem is linear. This approach has been used in [1] [3] [14].

In the present work formulation (14 - 23) was transformed into a single level formulation using the first approach, that is, by replacing the lower level problem by the optimality conditions given by its constraints, the constraints of its dual and the complementary slackness constraints, which were linearized using the *big-M* method. In that way, by replacing the lower level with its optimality conditions, variables which represent decisions of the users (v and w) are restricted to take values which solve problem (19 - 23). Therefore, the whole model will adjust the frequency values (variable y) so as to respect the constraints which are directly included in the upper level (among them, bus capacity) as well as the optimality conditions which represent the (uncapacitated) lower level problem.

After applying the KKT conditions, the resulting MILP model, equivalent to (14 - 23), is (24 - 51), where (33 - 36) correspond to the constraints of the dual of problem (19 - 23), π_{nk}, ν_{ak}, and μ_{ak} are the dual variables corresponding to constraints (20), (21), and (22), respectively, s^1_{ak} and s^2_{ak} are slack variables associated with inequality constraints (21) and (22), respectively, and t^1_{ak}, t^2_{ak} and t^3_{nk} are slack variables associated with the inequality constraints (33), (34) and (35), respectively. The complementary slackness conditions are linearized by applying the big-M method (37 - 46), obtaining in this manner a MILP single level formulation.

4 Experiments for a small-sized example

In order to illustrate the application of the bilevel model explained in section 3, we show in Figure 3 the small-sized case considered.

The numbers close to the arcs indicate their corresponding travel times. There are two OD pairs, such that $O_1 = 1$, $O_2 = 2$, $D_1 = D_2 = 3$ and $\delta_1 = \delta_2 = 5$. We consider values of fleet size $B = 10$, bus capacity $\omega = 1.0$ and the set of possible frequencies $\Theta = \{1.0, 2.5, 5.0, 7.0, 9.0\}$. The lines defined for this case are $l_1 = \{(1,2),(2,3)\}$ and $l_2 = \{(1,3)\}$, both having symmetrical forward and backward itineraries.

Table 1 shows the results of applying three different variants of formulation (24 - 51) to the example of Figure 3, where τ (calculated in (24)) is the total travel time of the optimal solution and β (calculated in (25)) its corresponding fleet size; it also shows the line capacity (as defined in expression (13)) and the critical flow of each line (defined as the flow of the arc v_a with maximum flow on the line). Even though the model has a large number of variables, due to the small size of the instance, the execution time is negligible.

Table 1 Impact of adding the bus capacity constraint.

Model	cap. l_1	critical flow l_1	cap. l_2	critical flow l_2	τ	β
uncapacitated	9.0	$9/10\delta_1 + \delta_2 = 9.5$	1.0	$1/10\delta_1 = 0.5$	4.8	≤ 10
cap. single-level	9.0	$8/10\delta_1 + \delta_2 = 9.0$	1.0	$2/10\delta_1 = 1.0$	5.3	≤ 10
cap. bilevel	9.0	$9/11.5\delta_1 + \delta_2 = 8.9$	2.5	$2.5/11.5\delta_2 = 1.1$	≤ 4.8	11.5

$$\min_{y,v,w} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{24}$$

$$\text{s.t.} \quad \sum_{l\in L}\sum_{f\in\Theta} \theta_f y_{lf} \sum_{a\in l} c_a \leq B, \tag{25}$$

$$\sum_{f\in\Theta} y_{lf} = 1 \qquad \forall\, l \in L, \tag{26}$$

$$\sum_{k\in K} v_{ak} \leq \sum_{f\in\Theta} y_{l(a)f}\theta_f \omega \qquad \forall\, a \in A^T, \tag{27}$$

$$\sum_{a\in A_n^+} v_{ak} - \sum_{a\in A_n^-} v_{ak} = b_{nk} \qquad \forall\, n \in N, k \in K, \tag{28}$$

$$v_{ak} \leq \theta_{f(a)} w_{nk} \qquad \forall\, a \in A_n^+, n \in N^P, k \in K, \tag{29}$$

$$v_{ak} \geq 0 \qquad \forall\, a \in A, k \in K, \tag{30}$$

$$v_{ak} \leq \delta_k y_{l(a)f(a)} \qquad \forall\, a \in A^B, k \in K, \tag{31}$$

$$y_{lf} \in \{0,1\} \qquad \forall\, l \in L, f \in \Theta, \tag{32}$$

$$\pi_{ik} - \pi_{jk} \leq c_a \qquad \forall\, a = (i,j) \in A - A^B, k \in K, \tag{33}$$

$$\pi_{ik} - \pi_{jk} - \mu_{ak} - \nu_{ak} \leq c_a \qquad \forall\, a = (i,j) \in A^B, k \in K, \tag{34}$$

$$\sum_{a\in A_n^{B+}} \theta_{f(a)}\nu_{ak} \leq 1 \qquad \forall\, n \in N, k \in K, \tag{35}$$

$$\mu_{ak}, \nu_{ak} \geq 0 \qquad \forall\, a \in A^B, k \in K, \tag{36}$$

$$\theta_{f(a)} w_{ik} - v_{ak} \leq s^1_{ak} M \qquad \forall\, a = (i,j) \in A^B, k \in K, \tag{37}$$

$$\nu_{ak} \leq (1 - s^1_{ak})M \qquad \forall\, a \in A^B, k \in K, \tag{38}$$

$$\delta_k y_{l(a)f(a)} - v_{ak} \leq s^2_{ak} M \qquad \forall\, a \in A^B, k \in K, \tag{39}$$

$$\mu_{ak} \leq (1 - s^2_{ak})M \qquad \forall\, a \in A^B, k \in K, \tag{40}$$

$$c_a - \pi_{ik} + \pi_{jk} \leq t^1_{ak} M \qquad \forall\, a = (i,j) \in A - A^B, k \in K, \tag{41}$$

$$v_{ak} \leq (1 - t^1_{ak})M \qquad \forall\, a \in A - A^B, k \in K, \tag{42}$$

$$c_a - \pi_{ik} + \pi_{jk} + \mu_{ak} + \nu_{ak} \leq t^2_{ak} M \qquad \forall\, a = (i,j) \in A^B, k \in K, \tag{43}$$

$$v_{ak} \leq (1 - t^2_{ak})M \qquad \forall\, a \in A^B, k \in K, \tag{44}$$

$$1 - \sum_{a\in A_n^{B+}} \theta_{f(a)}\nu_{ak} \leq t^3_{nk} M \qquad \forall\, n \in N, k \in K, \tag{45}$$

$$w_{nk} \leq (1 - t^3_{nk})M \qquad \forall\, n \in N, k \in K, \tag{46}$$

$$s^1_{ak} \in \{0,1\} \qquad \forall\, a \in A, k \in K, \tag{47}$$

$$s^2_{ak} \in \{0,1\} \qquad \forall\, a \in A^B, k \in K, \tag{48}$$

$$t^1_{ak} \in \{0,1\} \qquad \forall\, a \in A - A^B, k \in K, \tag{49}$$

$$t^2_{ak} \in \{0,1\} \qquad \forall\, a \in A^B, k \in K, \tag{50}$$

$$t^3_{nk} \in \{0,1\} \qquad \forall\, n \in N, k \in K \tag{51}$$

4.1 Experiment 1: comparison of uncapacitated and single level capacitated models

The first line of Table 1 shows the results of applying the uncapacitated model (5 - 12). When capacities are not considered, the entire flow of OD pair 2 uses l_1, while the flow of OD pair 1 is distributed between both lines (4.5 uses l_1 and 0.5 uses l_2) due to the flow splitting constraint (9).

When we consider bus capacities in the original uncapacitated model (second line of the table), adding the constraint directly, we obtain the same setting of frequencies but with a different assignment of flows. In this case, 1.0 units of the demand corresponding to OD pair 1 uses l_2. This is because l_1 has capacity to accommodate only up to 9.0 units of flow. The 0.5 units of flow corresponding to OD pair 1, which were moved from l_1 to l_2 represent a set of users who are forced to use a sub-optimal hyperpath, knowing the existence of a better one, that is, they behave in an unrealistic way. Moreover, we note that the model is not able to represent this situation consistently, since it can not represent different waiting times for passengers corresponding to the same OD pair at the same stop (variables w_{nk}).

The example shows through a numerical application, the consequences of solving the capacitated problem in a straightforward (not realistic) way. When we apply the bilevel model (24 - 51) to the same case, we obtain no feasible solution. This is due to the fleet size constraint, that does not allow for an increase of frequencies in order to accommodate the demand on the lines that the users want to use; moreover, the model is not able to change the frequencies in such a way as to redistribute the flows in order to respect the line capacities. That difficulty was already noted in [7]. In order to overcome this difficulty, we identify two approaches in the literature:

- Soften the bus capacity constraint, by moving it as a term of the objective function [7].
- Allow the model to increase the fleet size, by including its respective constraint in the objective function [19].

By adopting the first approach, the solutions obtained may violate the bus capacity constraint; the higher the violation, the less valid is the corresponding assignment of flows, which is done assuming sufficient capacity. On the other hand, the second approach assumes that the fleet size can be increased. This may be a reasonable assumption in the context of strategic planning, where the model can be used to estimate the investment required to offer a given level of service. In this case, by adding a new objective function the resulting model becomes multi-objective, which requires a special treatment depending on how this nature is represented: for example, by setting appropriate weights or calculating non-dominated solutions [21].

4.2 Experiment 2: calculation of required fleet size

Considering the discussion above, another possible application of the bilevel model to the capacitated case would be to state the fleet size minimization as upper level objective, subject to a constraint of maximum travel time; that is, swapping objective function (14) and constraint (15).

The results of applying this model to the small instance can be found in the third line of Table 1, where we state a maximum travel time equal to 4.8 (the optimal value of the uncapacitated model). The optimal value in this case (which corresponds to the fleet size), is equal to 11.5. The interpretation of the result is that in order to obtain a setting of

frequencies which respects the bus capacity constraint while at the same time producing a total travel time which is no worse than the one corresponding to the uncapacitated case, the fleet size should be increased in 15%.

5 Conclusions and further research

In the present work we propose a new bilevel formulation for transit frequency optimization, based on the model presented in [20]. The proposed model considers individual passenger route choice, using an assignment model [24], as well as considering the waiting time of the users and the bus capacity when measuring the performance of the system. We derived a mixed integer linear programming (MILP) formulation which is equivalent to the bilevel one, that is susceptible of being solved by common solvers using standard MILP techniques, for small-sized problem instances.

We have also explored the bilevel nature of the problem by applying different formulations to the same example instance. The results obtained suggest that a true bilevel approach should be considered whenever bus capacities are contemplated, and that uncapacitated models are able to produce solutions that are not appropriate in contexts where the transit system is operating over its capacity.

We note that all variants of the bilevel model discussed here maintain the DCLB structure. This enables to apply exact solution methods. However, the existing (general purpose) solution methods for this kind of bilevel problems [2] [3] [12] do not necessarily handle models with many variables and constraints, as it is the case of frequency optimization problems. Therefore, further research is needed in order to devise tailored solution methods for the specific problem. An example of such an approach can be found in [18]. Metaheuristic techniques may also aid in finding good solutions to solve the transit frequency optimization problem. The *Tabu Search* [17] based metaheuristic presented in [20] to solve a single level instance of the problem might also be extended to cope with a bilevel program. There are a growing number of metaheuristic approaches that deal with bilevel problems. A good survey can be found in [11].

It is also desirable to apply the proposed formulation to instances corresponding to real cities of medium size, in order to study the scalability of the method and the improvements obtained when compared to the current solutions of real transportation systems. The addition of other constraints, such as enforcing a maximum waiting time for users of the transit system may further help achieving solutions of good performance in real world contexts.

Regarding capacitated models, a formal criterion for switching between uncongested and congested frequency optimization models would be desirable to establish.

References

1 José M. Arroyo. Bilevel programming applied to power system vulnerability analysis under multiple contingencies. *IET Generation, Transmission and Distribution*, 4(2):178–190, 2010. doi:10.1049/iet-gtd.2009.0098.

2 Jonathan F. Bard. *Practical bilevel optimization*. Kluwer, 1998.

3 Luis Baringo and Antonio J. Conejo. Transmission and Wind Power Investment. *IEEE Transactions on Power Systems*, 27(2):885–893, 2012. doi:10.1109/TPWRS.2011.2170441.

4 Avishai Ceder and Nigel H. M. Wilson. Bus network design. *Transportation Research B*, 20(4):331–344, 1986.

5 Manuel Cepeda, Roberto Cominetti, and Michael Florian. A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. *Transportation Research B*, 40(6):437–459, 2006.

6 Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

7 Isabelle Constantin. *L'optimisation des fréquences d'un réseau de transport en commun*. Doctorate thesis on informatics, Université de Montréal, 1992.

8 Isabelle Constantin and Michael Florian. Optimizing frequencies in a transit network: a nonlinear bi-level programming approach. *International Transactions in Operational Research*, 2(2):149–164, 1995.

9 Joaquín de Cea and Enrique Fernández. Transit assignment for congested public transport systems: An equilibrium model. *Transportation Science*, 27(2):133–147, 1993.

10 Guy Desaulniers and Mark D. Hickman. Public transit. *Transportation*, 14:69–127, 2007. doi:10.1016/S0927-0507(06)14002-5.

11 El-Ghazali Talbi. Metaheuristics for Bi-level Optimization. *Studies in Computational Intelligence*, 482, 2013.

12 José Fortuny-Amat and Bruce McCarl. A representation and economic interpretation of a two-level programming problem. *Journal of the Operational Research Society*, 32:783–792, 1981.

13 Ziyou Gao, Huijun Sun, and Lian Long Shan. A continuous equilibrium network design model and algorithm for transit systems. *Transportation Research B*, 38(3):235–250, 2004.

14 Lina P. Garcés, Antonio J. Conejo, Raquel García-Bertrand, and Rubén Romero. A bilevel approach to transmission expansion planning within a market environment. *IEEE Transactions on Power Systems*, 24(3):1513–1522, 2009. doi:10.1109/TPWRS.2009.2021230.

15 Guido Gentile, Michael Florian, Younes Hamdouch, Oded Cats, and Agostino Nuzzolo. *The Theory of Transit Assignment: Basic Modelling Frameworks*, pages 287–386. Springer International Publishing, Cham, 2016. doi:10.1007/978-3-319-25082-3_6.

16 Ricardo Giesen, Héctor Martínez, Antonio Mauttone, and María E. Urquhart. A method for solving the multi-objective transit frequency optimization problem. *Journal of Advanced Transportation*, 50, 2017. doi:10.1002/atr.1461.

17 Fred Glover. Tabu search part I. *ORSA Journal on Computing*, 1(3):190–206, 1989.

18 Marc Goerigk and Marie Schmidt. Line planning with user-optimal route choice. *European Journal of Operational Research*, 259:424–436, 2017.

19 Carola Leiva, Juan Carlos Muñoz, Ricardo Giesen, and Homero Larrain. Design of limited-stop services for an urban bus corridor with capacity constraints. *Transportation Research B*, 44(10):1186–1201, 2010.

20 Héctor Martínez, Antonio Mauttone, and María E. Urquhart. Frequency optimization in public transportation systems: formulations and metaheuristic approach. *European Journal of Operational Research*, 236(1):27–36, 2014. doi:10.1016/j.ejor.2013.11.007.

21 Antonio Mauttone and María E. Urquhart. A multi-objective metaheuristic approach for the transit network design problem. *Public Transport*, 1(4):253–273, 2009.

22 Sang Nguyen and Stefano Pallottino. Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research*, 37(2):176–186, 1988.

23 Francisco Ruisánchez, Luigi dell'Olio, and Angel Ibeas. Design of a tabu search algorithm for assigning optimal bus sizes and frequencies in urban transport services. *Journal of Advanced Transportation*, 46(4):366–377, 2012.

24 Heinz Spiess and Michael Florian. Optimal strategies: a new assignment model for transit networks. *Transportation Research B*, 23(2):83–102, 1989.