PEDECIBA Informática

Instituto de Computación, Facultad de Ingeniería

Universidad de la República

Montevideo, Uruguay

TESIS DE MAESTRÍA en Informática

Frequency optimization in public transportation with strict capacity constraints. A bilevel programming approach.

Agustín Arizti

agustin.arizti@gmail.com

Octubre de 2024

Supervisor: Antonio Mauttone

Orientadores de tesis: Antonio Mauttone y María E. Urquhart

Frequency optimization in public transportation
with strict capacity constraints.
A bilevel programming approach.
Arizti, Agustín
ISSN X
Tesis de Maestría en Informática
Reporte Técnico
PEDECIBA
Instituto de Computación - Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay, Octubre de 2024

Resumen

En la presente tesis se considera el problema de optimización de frecuencias en sistemas de transporte público basados en ómnibus. El objetivo del problema es determinar el intervalo de tiempo entre pasadas de ómnibus consecutivos para un conjunto de líneas. Las soluciones deben satisfacer una demanda origen-destino dada y a su vez considerar el interés tanto de los usuarios como de los operadores, en un contexto de restricciones asociadas a la infraestructura subyacente, presupuesto y nivel de servicio ofrecido.

A efectos de considerar eventos de congestión sobre las líneas de transporte (esto es, cuando las líneas operan al límite de su capacidad en relación a la demanda que atraen), se extiende un modelo existente agregando una restricción de capacidad en los ómnibus, utilizando un sub-modelo que asigna la demanda de forma de respetar la decisión de los usuarios en cuanto a la elección de las distintas líneas. La formulación obtenida de esta forma es binivel, y mediante una reformulación del segundo nivel usando las condiciones de optimalidad, es convertida a una formulación lineal entera mixta (MILP) que puede resolverse de forma óptima para instancias de pequeñas dimensiones utilizando técnicas del estado del arte empleadas para los MILP.

Para estudiar el comportamiento del modelo propuesto se exploran diversas variantes de esta última formulación. Con el propósito de comparar distintas soluciones, y evaluar la factibilidad de implementarlas en escenarios reales, se aplica el modelo propuesto a un caso de pequeñas dimensiones, así como a un caso real que ha sido estudiado en otras publicaciones. A efectos de brindar un mejor nivel de servicio a los usuarios del sistema, se estudian los efectos de añadir a la formulación una restricción de tiempo de espera máximo de los usuarios en las paradas.

Se concluye que un enfoque binivel es necesario toda vez que se consideren capacidades en los ómnibus. Esto supone la necesidad de contar con modelos que incorporen tanto el comportamiento de los usuarios, como capacidades y el tiempo de espera en las paradas. La inclusión simultánea de todos estos aspectos en una formulación de programación matemática constituye una novedad en la literatura. A su vez, mediante un caso de estudio correspondiente a una ciudad real, se exploran algunas de las dificultades que surgen al tener en cuenta capacidades y distintos tamaños de flota. Gracias al estudio de distintas medidas, como las capacidades de las líneas, y tiempos máximos de espera en las paradas, con la ayuda de una inspección visual, es posible identificar los sectores más problemáticos de una red de transporte, permitiendo de esta forma una discusión detallada de los desafíos que pueden surgir en contextos reales de operación y contribuyendo de esta manera en el diseño de soluciones alternativas.

Finalmente, se analiza el límite en cuanto al tamaño de instancias susceptibles de ser resueltas de forma exacta utilizando el modelo propuesto.

Palabras clave: Transporte, Transporte público con capacidades, Optimización de frecuencias, Programación lineal entera mixta, Programación binivel.

Abstract

In this thesis, we consider the problem of frequency optimization in public transit systems based on buses. The objective of the problem is to determine the time interval between subsequent buses for a set of transportation lines. The solutions should satisfy a given origin-destination demand while considering the interests of users and operators in the context of constraints pertaining to infrastructure, budget, and service performance.

To consider congestion on the transportation lines (that is, when the lines operate at the limit of their capacities in relation to the attracted demand), we extend an existing model by adding a constraint on bus capacities while respecting user choice on the lines that can drive users to their destinations. The resulting formulation is bilevel and is then transformed, by means of applying optimality conditions on the lower level, into a mixed integer linear programming formulation (MILP) that can be solved to optimality over small instances using state-of-the-art MILP techniques.

To study the nature of the model, we analyze different variants of the proposed bilevel formulation. Furthermore, we compare solutions and evaluate their feasibility of being applied to real-world scenarios. In order to do that, we apply the model to a small test case and to a real one published in the literature. To provide a better level of service to the users of the system, we study the effects of adding to the proposed formulation a constraint on the maximum waiting times of the users allowed at the bus stops.

We conclude that a bilevel approach should be considered whenever bus capacities are contemplated; thus, there is a need for models that incorporate the behavior of the users, the waiting times at the stops, and bus capacities. To the best of our knowledge, the simultaneous inclusion of all of the aforementioned aspects in a single mathematical programming formulation has not been studied. Moreover, by using a test case corresponding to an actual city, we explore some underlying issues that arise whenever bus capacities and different fleet sizes are considered. By studying measures such as line capacities and maximum waiting times, with the aid of visual inspection, problematic sectors of the line network can be quickly identified. This allows a more thorough discussion of the issues that could arise in real-life contexts and helps devise alternative solutions.

Finally, we analyze the limit regarding the size of the instances that can be resolved by the proposed model.

Keywords: Transportation, Public transport capacity, Transit frequency optimization, Mixed integer linear programming, Bilevel programming.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deep gratitude to my supervisors, Antonio Mauttone, and María E. Urquhart, for all their support, guidance, and help throughout the research and writing of this thesis.

A special thanks to my mother, Graciela, for her encouragement and support in all my academic endeavors, and to my wife, Paulina, and our children, Valentín and Ainhoa, for their patience, love, and support.

Chapter 1

Introduction

1.1 Problem introduction

Public transportation systems are essential to most cities of the world, enabling their inhabitants to travel along the different parts of a city. This thesis is about models for the optimal assignment of frequencies to lines in such urban public transportation systems.

The main actors that participate in a public transportation system are [14]:

- Users: the passengers of the bus lines, that is, any person with specific transport needs who can use the system.
- Operators: the companies that provide the transport services, including the economic resources for the operation of vehicles, fuel, crew, and general maintenance.

The design of a public transportation system needs to consider monetary costs, which range from fixed costs due to the construction of the infrastructure to variable costs due to the operation of the services [57]. Generally, governments are in charge of the construction of the underlying infrastructure, while the operators (usually private companies) provide the services (lines), and users pay a given fare to use the transportation system [25].

As well as monetary costs, public transportation systems must also consider the users' interest, e.g., in providing reasonable travel times, waiting times, and number of transfers. The frequency setting directly affects both concerns, influencing the level of service offered to the users (waiting time, the capacity of the lines) and the costs that planners need to incur to run the system, since the frequency of the lines determines the fleet size [25].

Due to the complexity of the design of public transportation systems, which involves several different decisions and conflicting objectives, the problem is usually approached in a series of steps. In that sense, a division of planning stages, defined according to the planning horizon of each, is helpful [14]. There are different stages for the design of a public transportation system based on buses. Assuming that the underlying infrastructure (street networks, stops) is already available, the literature identifies five stages [14] that are usually performed sequentially in real systems:

- Route network design, the number of lines, and their topology.

- Frequency setting, where the problem is to determine the time interval between subsequent buses of each line.
- Timetable design, to decide on the departure and arrival time of each bus in every line.
- Fleet assignment, where considering constraints of available fleet and depot location, the problem is to decide the itinerary of trips that each bus must take.
- Crew assignment, where, subject to contraints of working rules, drivers and possibly other staff are assigned to operate each bus.

The decisions made at each stage influence the decisions at later stages, and they are approached considering different planning horizons, depending on whether the context of the planning is strategic (long term), tactical (medium term), or operational (short term). Frequency setting decisions are usually part of tactical planning [25], even though an initial frequency setting is necessary to evaluate the decisions made during route network design, which happens on a strategic basis.

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) [43]. Therefore, to measure a transportation system's performance from the users' viewpoint, models should consider how users behave when faced with the choice of a specific bus line from a set of candidate lines that can drive them to their destinations. Such is the responsibility of an *assignment sub-model*, a model that, by applying a set of hypotheses on how users behave, selects the appropriate lines in order to satisfy travel demands. The assignment model is in itself an optimization problem, usually having mathematical formulations which are difficult to address with effective and efficient solution methods, particularly when the influence of bus capacity is considered in the modelling of user behavior [15]. Therefore, the difficulty of the overall frequency optimization model is strongly determined by the complexity of the underlying assignment sub-model.

Due to the challenges described above, some proposals in the literature avoid to consider bus capacities at the same time as user behavior, which often results in models that do not represent appropriately situations where the bus capacity constraint is an issue. When considering bus capacity constraints, existing models might dismiss user behavior: in those cases, planners freely assign demand to the different lines in order to minimize the total overall travel time of the users, among other objectives and constraints. This assignment is usually unrealistic when bus capacities are considered because it expects users to behave in ways conflicting with their personal interests. Then again, when user behavior is represented through an assignment sub-model, bus capacities are usually avoided in the context of frequency optimization. Moreover, different assignment sub-models exhibit different degrees of realism reflected by the hypothesis considered and the context in which they are applied to [37].

Including the bus capacity constraint alongside an assignment sub-model changes the nature of frequency optimization, turning an uncapacitated single-level formulation into a bilevel one [7]. In bilevel optimization problems, exactly two decision-makers exist, and their objectives do not necessarily coincide. Furthermore, the individual decisions each one can make influence the decisions of the other. A sequential order is imposed, in which the

agent that comes first is called the leader, while the agent that reacts to the leader's actions is called the follower. It is important to stress that, even though the individual decisions of each agent can affect the other, there are no cooperation mechanisms whatsoever: each agent makes its decision in regard to its individual benefit (objective function).

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), as can be seen in [7, 18].

The bilevel nature of the frequency optimization problem is explained by 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. This topic will be further explained in the following chapters.

This thesis assumes the following hypothesis, which limits the scope of the proposed models:

- We do not consider other modes of public transportation (i.e., train systems) or any private means of transportation such as private cars.
- The set of users is fixed, without other alternatives for traveling (captive clients, inelastic demand).
- The fare charged for using the service is not considered, which means users are oblivious to fares when using the system.
- Advanced traveler information systems (ATIS) are not taken into account, so user behavior is considered independent of such assistance.
- In transportation systems that present certain characteristics (high frequency, special infrastructure, etc.), the waiting time and transfers are not perceived as negatively from the users' point of view. This work assumes that users are sensitive to waiting times. Transfers are considered implicitly by means of the waiting time for the next line; they are not penalized explicitly.

The remainder of this thesis is organized as follows. In Section 1.2, we present a review of related literature and the contributions of this work. The proposed mathematical programming formulation is presented in Chapter 3. In Chapter 4, we show computational experiments using different test cases. We begin by applying the model to a simple test case, exploring alternative formulations and the bilevel nature of the problem. We then tackle an instance of a larger size used in the literature and for which solutions have been reported. We also conduct computational experiments over this instance to explore the model's sensitivity to certain parameters. Finally, we explore the scalability of the proposed formulation by trying to solve a bigger case corresponding to an actual city. We conclude the work and refer to future research directions in Chapter 5.

1.2 Related literature and statement of contribution

1.2.1 Literature review

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

The majority of the models in frequency optimization represent the demand between different zones of a city by means of an origin-destination (OD) matrix, where each positive element is called an *OD pair*. Generally, the matrix indices represent special nodes called *centroids* that act as a proxy of the different zones of a city [23]. A given origin-destination matrix expresses the number of trips per person that should be satisfied between nodes in a specific period of the day.

The infrastructure (street and bus stops) is typically formulated in terms of a graph, with nodes representing bus stops, centroids, and street endpoints. Arcs usually represent segments of a line's path or specific events or actions like waiting for a given line or performing a transfer. In some models, walking arcs are incorporated, which typically connect a centroid node to a bus stop (and vice versa), while some models merge both concepts and assume that the demand is originated at the bus stops [23].

Several works related to models and algorithms for frequency optimization in public transportation systems present approximate methods, based on heuristics or metaheuristics [12] without considering explicit formulations. In some cases, an explicit formulation is proposed (that is, a formulation defined by mathematical expressions representing decision variables, constraints, and objective functions), but where the assignment sub-model that represents user behavior is not explicit, forcing users to behave as to optimize some global optimum according to the objectives considered [65].

Since frequencies impact the total travel time of users of the system, some works consider both route design and frequency assignment in a single problem, as at least an initial frequency setting is required in order to evaluate a set of routes [40, 65]. Moreover, frequencies determine whether the transportation system has enough capacity. More often, the problem is decomposed into subproblems, first solving the design of the routes and, later, setting frequencies for each of the lines.

Due to the interaction among the different lines and since the waiting time is inversely proportional to the frequencies, nonlinear models are frequent in the literature. When included, existing models consider assignment sub-models with varying degrees of realism. The size of the test cases considered also varies, from artificial and small-sized instances to medium-sized cases representing real cities comprising up to 100 lines, approximately [76].

In [65], an explicit formulation is defined for the line planning problem in public transport that models both line design and frequency setting. A column-generation approach is proposed, and the authors demonstrate that their proposed model is NP-Hard. The waiting time of users and transfers are not considered. The behavior of the users is not explicitly stated, so they will choose lines in order to minimize a global optimum that considers the travel time of all of the users of the system and, at the same time, the cost of operation of the services.

In [70], the main concepts related to route planning and frequency setting are presented, as well as many of the mathematical models and solution methods proposed in the literature up to the date of publication.

Frequency optimization is also studied in [69]. The problem is formulated as a nonlinear programming in the form of a compound minimization problem. It attempts to minimize the users' walking, on-board, and waiting times. It also considers bus capacities as a constraint. There is no explicit assignment sub-model; the demand is divided among the different lines according to entropy and bus capacities. An iterative algorithm to solve the problem is proposed. The algorithm is tested with a case study representing the bus network in the town of Linköping (Sweden).

In [44], the problem addressed is that of allocating a fleet of buses among routes in networks characterized by having extensively overlapping routes and where buses frequently operate at, or close to, capacity. Since the intention is to alleviate congestion on heavily used lines, the objective function states the minimization of the occupancy level at the most heavily loaded point on any route in the system. Bus capacities and user behavior are considered, albeit the latter in a non-explicit manner by encoding a set of preferences and rules that users follow. A heuristic method is proposed, and a small case comprising 6 nodes and 3 routes is reported.

In [20], a nonlinear bilevel formulation for frequency optimization is proposed. It incorporates an explicit assignment model [72] 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 diminishing the monetary costs by constraining the fleet size. Both levels have the same objective function: minimizing the overall travel time. 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 multi-objective model is proposed in [33], where the objectives are to minimize the users' overall travel time and the planners' cost (using line frequencies as a proxy). The model also incorporates congestion in the behavior of the users through an assignment model that calculates effective frequencies [22], that is, the real frequencies that users experience when, due to lack of capacity, they can not board the desired lines. The resolution method consists of a sensitivity analysis procedure, and a single test case of minimal size (4 nodes and 4 lines) is presented.

A genetic algorithm for bus frequency optimization is proposed in [76], where the authors state a bilevel programming model that aims to minimize the total travel time of passengers subject to a constraint on the overall fleet size. In this case, the behavior of the users is explicit by incorporating the optimal strategies [72] assignment model. In order to solve the bilevel model, an iterative approach consisting of a genetic algorithm and a label-marking method is proposed. The model and the algorithms are illustrated with small and medium-sized test cases.

A bilevel model is proposed in [68], where the upper level seeks to improve an overall cost function and the lower level consists of the capacity constrained assignment problem formulated in [22]. Tabu Search [39] is used as the heuristic resolution method.

In [55], a MILP formulation is proposed that models user behavior using the optimal strategies [72] 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 using a commercial solver on small instances; a metaheuristic based on Tabu Search is used for larger instances. The metaheuristic approach is tested using real case studies corresponding to the cities of Rivera, Uruguay (84 nodes, 143 edges, and 13 lines) and Montevideo, Uruguay (4945 nodes, 14672 edges, and 133 lines).

More recently, [40] proposed 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 decision variables, but only to incorporate bus capacities into the model, ignoring the waiting time of the users of the system. Exact solution methods are proposed, and a genetic algorithm is used in order to solve large-scale instances. The authors also explore the differences of considering user behavior as part of the model (i.e., letting users choose the lines) as opposed to implicitly letting planners assign the demand to the different lines.

We can conclude that most frequency optimization models proposed in the literature have similar characteristics. There are variations on whether models consider explicit assignment sub-models and, when they do, which hypotheses are assumed. All solution methods but the ones in [40] and [55] are approximated, driven by heuristics and metaheuristics.

1.2.2 Motivation of this thesis

The primary motivation of this thesis is the study of the problem of frequency optimization in public transportation systems. In particular, we are interested in contexts where the capacity of the buses is relevant, that is, when passenger affluence causes the services to operate at the limit of their capacities.

The aim is to improve the realism of existing approaches by considering several important aspects of the problem that have not been considered simultaneously, that is, passenger behavior, bus capacity constraints, and the waiting time of the users.

It is also of interest to include the aforementioned characteristics in a model suitable to be formulated explicitly, in order to obtain solutions with optimal guarantees, since most of the literature presented in Subsection 1.2.1 is focused on obtaining approximate solutions by heuristic or metaheuristic methods.

In doing so, several interesting challenges arise:

- The existing mathematical programming formulations fail to simultaneously include several relevant aspects, such as passenger behavior, bus capacity constraints, and the waiting time of the users of the system.
- Considering the influence of capacities over user behavior at the same time as the waiting time, changes the nature of the problem from a single-level one to a bilevel problem. Bilevel programming problems occur naturally in various real-world scenarios [9, 18, 47, 54] but are nevertheless hard to solve [32, 45, 59].
- Most existing methods are based on heuristics that are not driven by explicit formulations. The proposed methods then cannot give an optimality gap for the reported results, which hinders the ability to judge the quality of the reported solutions.

- Frequency optimization is a relevant problem in the world, having a direct impact to society. Public transportation systems are essential to most cities of the world, and frequency setting is a concern both during strategic planning, when the itinerary of the lines is designed, as well as during tactical planning, whenever the frequency of the lines requires adjustments due to seasonal variations [25]. Moreover, the setting of frequencies significantly impacts the acceptance of a public transportation system from a customer's point of view, having a great influence in the perceived quality of its users.
- Frequency optimization is a difficult problem, due to nonlinear and combinatorial characteristics and the modeling of user behavior by means of an assignment submodel (a complex problem in itself).

Some interesting aspects arise when considering the bilevel nature of the problem. Is it indispensable to employ a bilevel approach when seeking solutions? And how do the bilevel nature and the interplay of operators and users affect the solutions encountered? In addition, we study the flexibility of such a formulation in terms of adding new constraints and exploring other aspects of reality, such as ensuring that individual users do not experience arbitrarily high waiting times at bus stops. Finally, we examine whether the proposed formulation is suitable for use in a real-life context as part of the decision-making process in determining frequencies.

1.2.3 Contributions of this thesis

The contributions of this thesis are:

- Leverage the advances in the fields of bilevel optimization and mixed integer linear programming (MILP) for solving the frequency setting problem in public transport, using exact methods able to ensure enough capacity over the bus lines.
- Improve the realism of existing models by considering passenger behavior as well as the bus capacity constraint and the waiting time of the users, which, under several scenarios, is the most relevant part of the total travel time. To the best of our knowledge, including all of the aforementioned aspects simultaneously in a single formulation has yet to be studied (for example, in [40], the waiting time is not considered).
- Propose a bilevel formulation that is converted to a MILP formulation suitable of being solved exactly by using state-of-the-art solvers. In this manner, we can obtain optimal solutions while other related works (such as [20]) are only suitable to be solved approximately using heuristic or metaheuristic procedures.
- Analyze the impact of adding the bus capacity constraint to the model and study alternative formulations to better understand the problem's nature. Is it possible to ensure capacity with the proposed approach?
- Apply the formulation to a test case published in the literature in order to compare the solutions and analyze the feasibility of the method. Attempt to solve a bigger instance based on an actual city to evaluate the proposed approach's scalability.

We also note that the work done in the context of this thesis resulted in the two following publications:

- Agustín Arizti, Antonio Mauttone, and María E. Urquhart. A Bilevel Approach to Frequency Optimization in Public Transportation Systems. In Ralf Borndörfer and Sabine Storandt, editors, 18th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2018), volume 65 of Open Access Series in Informatics (OASIcs), pages 7:1-7:13, Dagstuhl, Germany, 2018. Schloss Dagstuhl - Leibniz-Zentrum für Informatik. Presented by Arizti during the ATMOS conference in Helsinki, Finland, 2018.
- Antonio Mauttone, Agustín Arizti, and María E. Urquhart. Frequency optimization in public transportation with strict capacity constraints. Invited abstract in session WB-55: Public transportation, stream Transportation, EURO 2024 Copenhagen. Presented by Mauttone during the EURO conference in Copenhagen, Denmark, 2024.

Chapter 2

Background

This chapter presents several necessary concepts when studying the frequency optimization problem and the mathematical programming formulation proposed in this work.

In Section 2.1, we introduce the principal concepts of duality theory in linear programming, which will play an important role when reformulating the bilevel model into a single-level one. To better understand the proposed formulation, some familiarity with bilevel programming is required, and the basic theory is thus presented in Section 2.2. Even though the proposed formulation has a single objective, we introduce in Section 2.3 various concepts related to multi-objective optimization that will prove useful when analyzing the interplay between fleet sizes and total travel times in the solutions reported by our method. In Section 2.4, we present the concept of assignment model in the context of public transport, a critical component of any model of frequency optimization. Finally, in Section 2.5, we consider assignment under the circumstances of congestion in a public transit network, an important challenge in models of frequency optimization that consider bus capacities.

2.1 Duality in linear programming

Duality theory deals with the relation between a linear programming problem, commonly referred to as the "primal" problem, and another linear programming problem, commonly known as the "dual". This relationship has various practical and theoretical applications and will be of particular importance in the context of this thesis when transforming a bilevel formulation for the problem of frequency optimization into a single-level formulation. Most of this section's content has been adapted from [10].

One approach used in calculus to minimize a function subject to some equality constraints is the Lagrange multiplier method. The idea of the method is to avoid enforcing hard constraints by instead associating each constraint with a Lagrange multiplier, or price, p, with the amount by which it is violated in each restriction. When p is properly chosen, the violation is 0, and thus, the optimal solution to the constrained problem is also optimal for the unconstrained problem.

The same idea applies to linear programming. It is possible to associate a price variable with each constraint and then search for the prices that do not affect the optimal cost.

Consider a primal problem in standard form:

$$\min_{x} \quad cx \tag{2.1}$$
$$\text{s.t.} \quad Ax = b \tag{2.2}$$
$$x \geq 0 \tag{2.3}$$

and let x^* be an existing optimal solution. To work with an unconstrained problem, we can introduce a price vector p, to obtain the following relaxation of the primal problem:

$$\min_{x} \quad cx + p(b - Ax) \tag{2.4}$$
$$\text{s.t.} \quad x \geq 0 \tag{2.5}$$

where constraint $Ax = b$ is replaced by a penalty in the objective function. Since problem (2.4 - 2.5) is a relaxation of (2.1 - 2.3), the optimal cost of the relaxed problem $g(p)$ as a function of p can not be larger than the optimal primal cost cx^*. Formally:

$$g(p) = \min_{x \geq 0} \quad [cx + p(b - Ax)] \leq cx^* + p(b - Ax^*) = cx^* \tag{2.6}$$

The last equality follows because x^* is a feasible solution, thus satisfying $Ax^* = b$. Problem (2.4 - 2.5) results, for each p, in a lower bound of the primal problem. The tightest possible lower bound can be obtained by solving the following problem:

$$\max \quad g(p) \tag{2.7}$$

which is known as the *dual* problem. The main theorem in duality theory is that of *strong duality*, which states that if a linear programming problem has an optimal solution, so does its dual, and the respective optimal costs are equal. Thus, the tightest lower bound obtained by solving problem (2.7) is not just a lower bound but equal to the optimal cost cx^* of the primal problem.

The dual problem can be stated more explicitly. Using the definition of $g(p)$:

$$g(p) = \min_{x \geq 0} \quad [cx + p(b - Ax)] = pb + \min_{x \geq 0} \quad (c - pA)x \tag{2.8}$$

Noting that:

$$\min_{x \geq 0} \quad (c - pA)x = \begin{cases} 0 & \text{if } c - pA \geq 0, \\ -\infty & \text{otherwise} \end{cases} \tag{2.9}$$

Since the goal is to maximize $g(p)$, we can discard values of p where $g(p)$ equals $-\infty$. The dual problem can then be stated as the following linear programming problem:

$$\max_{p} \quad pb \tag{2.10}$$

$$\text{s.t.} \quad pA \leq c \tag{2.11}$$

Analogous derivations can be made for the cases where the x vector is free rather than constrained.

2.1.1 Main results

Several important theorems in linear programming result from duality theory. Earlier, we noted that the cost $g(p)$ of any dual solution is a lower bound of the primal optimal cost. Weak duality proves this result in a more general way.

Theorem 1 (Weak duality) *If x is a feasible solution to the primal problem and p is a feasible solution to the dual problem, then $pb \leq cx$.*

Which naturally leads to the following corollaries.

Corollary 1.1 *If the optimal cost in the primal is -∞, then the dual problem must be infeasible. If the optimal cost in the dual is +∞, then the primal problem must be infeasible.*

Corollary 1.2 *Let x and p be feasible solutions to the primal and the dual, respectively, and suppose that $pb = cx$. Then, x and p are optimal solutions to the primal and the dual, respectively.*

The main result in duality theory is probably that of *strong duality.*

Theorem 2 (Strong duality) *If a linear programming problem has an optimal solution, so does its dual, and the respective optimal costs are equal.*

Finally, an important relation between primal and dual optimal solutions is provided by the *complementary slackness* conditions.

Theorem 3 (Complementary slackness) *Let x and p be feasible solutions to the primal and the dual problem, respectively. The vectors x and p are optimal solutions for the two respective problems if and only if:*

$$p_i(a_i x - b_i) = 0 \qquad \forall\ i$$

$$(c_j - pA_j)x_j = 0 \qquad \forall\ j$$

2.1.2 Karush-Kuhn-Tucker conditions

Duality theory is closely related to the so-called *Karush-Kuhn-Tucker conditions* (hereafter, KKT conditions). The KKT conditions are necessary conditions that a local optimum must satisfy, providing some regularity conditions. The conditions are not exclusive to linear programming, but they also apply to nonlinear programming: it is sufficient that the objective function and the constraint functions are all differentiable. When additional conditions hold, such as when the minimization problem considered is convex, they also constitute sufficient conditions for proving that a solution is a global optimum.

Consider the following problem:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 && \forall\ i = 1, ..., m \\ & h_i(x) = 0 && \forall\ i = 1, ..., p \end{aligned}$$

where all functions are differentiable. A point $x \in R^n$, alongside multipliers $u \in R^m$ and $v \in R^p$, verifies the KKT conditions for the problem above if:

$$\nabla f(x) + \sum_{i=1}^{m} u_i \nabla g_i(x) + \sum_{i=1}^{p} v_i \nabla h_i(x) = 0 \tag{2.12}$$

$$g_i(x) \leq 0, \qquad \forall\ i = 1, ..., m \tag{2.13}$$

$$h_i(x) = 0, \qquad \forall\ i = 1, ..., p \tag{2.14}$$

$$u_i \geq 0, \qquad \forall\ i = 1, ..., m \tag{2.15}$$

$$u_i g_i(x) = 0, \qquad \forall\ i = 1, ..., m \tag{2.16}$$

We note that u_i and v_i are the dual variables associated with constraints $g_i(x)$ and $h_i(x)$, respectively, while equality (2.16) is the complementary slackness condition.

Besides being a powerful theory that provides new geometric insights, duality theory has various applications, such as in the dual simplex method for solving linear programming. The KKT conditions applied to linear programming imply establishing primal and dual feasibility and complementary slackness. In this thesis context, the complementary slackness condition will prove helpful in Subsection 3.4.2 to reformulate a bilevel mathematical program into a single-level one.

2.2 Bilevel mathematical programming

The goal of bilevel optimization is to understand and solve real-life problems where two decision-makers exist who constitute a hierarchy. Furthermore, the objectives of these decision-makers do not necessarily coincide, and the individual decisions each one can make influence the decisions of the other. A bilevel program then always encompasses two problems, one for each of the decision-makers or agents. A sequential ordering is imposed, in which the agent that comes first is called the leader, while the agent that reacts to the leader's actions is called the follower. An *optimality constraint* in the upper-level problem

(the leader problem) establishes that one or several decision variables must be part of the optimal solution of yet another optimization problem (the follower problem, also known as the lower-level problem). In that way, the agent corresponding to the upper level of the hierarchy makes decisions that constrain the decisions of the agent corresponding to the lower level and needs to anticipate the reaction of the lower level.

The general bilevel programming problem (from now on, BLPP) is formulated as follows [7]:

$$\min_{x \in X, y} \quad F(x, y) \tag{2.17}$$

$$\text{s.t.} \quad G(x, y) \leq 0 \tag{2.18}$$

$$y \in arg \min_{y' \in Y} \quad f(x, y') \tag{2.19}$$

$$\text{s.t.} \quad g(x, y') \leq 0 \tag{2.20}$$

Where the leader has control over the vectors $x \in X \subseteq R^n$ and the follower controls the vectors $y, y' \in Y \subseteq R^m$, $F, f : R^n \times R^m \to R$, $G : R^n \times R^m \to R^p$, $g : R^n \times R^m \to R^q$. All the functions are assumed to be continuous and twice differentiable. The sets X and Y may impose additional restrictions, such as nonnegativity or integrality.

Hereafter, we adopt the following simplified BLPP notation, where the reference to *argmin* and the distinction between variables y and y' is removed:

$$\min_{x \in X, y} \quad F(x, y) \tag{2.21}$$

$$\text{s.t.} \quad G(x, y) \leq 0 \tag{2.22}$$

$$\min_{y \in Y} \quad f(x, y) \tag{2.23}$$

$$\text{s.t.} \quad g(x, y) \leq 0 \tag{2.24}$$

The decision-making process modeled by a bilevel mathematical program can be seen as carried out in two sequential stages [31]. From the point of view of the real problem under consideration, the leader should make a decision first by selecting appropriate values over x in order to minimize $F(x, y)$, subject to additional constraints, as denoted by $G(x, y)$. Notice that this does not suggest a mandatory ordering of steps in the design of resolution methods but only illustrates the scenario from the point of view of the decision-makers: resolution methods exist, such as co-evolutionary algorithms [51], that operate on both problems in parallel. When the leader made a decision, no information is provided regarding the values that the vector y will take because those values represent solely the decision made by the follower. So, in this first instance, the leader chooses values for vector x, and then the follower, using the input on x given by the leader, reacts accordingly by selecting the vector y in such a way as to minimize its own objective function $f(x, y)$, subject to its own constraints $g(x, y)$. Notice that, for the follower, the vector x represents constant values rather than variables. After the follower makes a decision that is optimal at the lower level, the upper-level problem must also enforce feasibility since some constraints might depend on the values given by the follower through vector y.

Bilevel programming problems are hard to solve. Complexity results in [32] have shown that even the linear version of the BLPP (where both the upper and lower level problems are linear) is NP-hard, and afterward, these results were strengthened in [45] where it was demonstrated that the linear BLPP is, in fact, strongly NP-hard. Checking global or local optimality is also NP-hard [59].

Many real-world problems can be modeled as bilevel optimization problems. Although the nature of several problems would be best represented by using a bilevel approach, since they are usually hard to solve, this is generally avoided, or the problem is reformulated in such a way that the bilevel nature does not hold anymore, for example, by dropping problematic constraints or considering simpler hypotheses. Observe that a trivial relaxation of the BLPP can be stated as follows:

$$\min_{x \in X, y} \quad F(x, y) \tag{2.25}$$

$$\text{s.t.} \quad G(x, y) \leq 0 \tag{2.26}$$

$$g(x, y) \leq 0 \tag{2.27}$$

Where the optimality constraint has been dropped, and the resulting formulation is single level.

Several properties that arise in bilevel programming problems may seem curious or counter-intuitive compared to the more traditional case of single-level mathematical programming. Even if the set conformed by all the solutions that satisfy constraints G and g is nonempty and compact, the existence of solutions for the BLPP is not guaranteed [7]. This can happen when, for a given fixed decision x^* by the leader, the follower has several optimal values of vector y to choose from, all yielding the same objective value. If x^* is part of the optimal solution, there is no guarantee (nor incentive) from the follower's point of view to choose the specific optimal y^* from the set of possible optimal solutions that would yield the optimal global value (x^*, y^*) of the BLPP. Different approaches exist that try to circumvent this issue by establishing mechanisms where the follower either cooperates with the leader or actively opposes him [24]. Another deviation from traditional single-level optimization problems is that in the BLPP case, the addition of irrelevant constraints (that is, constraints that are not active) affects the set of optimal solutions that can be obtained, so cautious consideration has to be paid when interpreting whether some restriction is necessary or useful [7].

Despite the computational complexity of the BLPP, several resolution methods exist, and some of them guarantee finding the global optimum. Some of the most popular exact methods to solve the various instances of the BLPP use the Karush Kuhn Tucker (KKT) or primal-dual conditions to reformulate the problem and convert it to a single-level equivalent (this is explained with more detail in Subsection 3.4.2). There are also specialized branch and bound algorithms [6], extreme point ranking procedures [11], and other alternative reformulations [8]. Approximate algorithms include heuristic and metaheuristic approaches, such as those presented in [30], where a transformation strategy from a BLPP to a multi-objective optimization problem is proposed, co-evolutionary approaches [51], and more traditional metaheuristic approaches (ranging from genetic algorithms [64], Greedy Randomized Adaptive Search Procedures (GRASP) [67], Tabu Search

[36] or Particle Swarm Optimization [48], among others) that tackle some of the aforementioned single-level reformulations that can be achieved by using, for example, the KKT approach.

Bilevel optimization problems appear in areas as diverse as transportation [20], production [73], energy [9] and environmental protection [73].

2.3 Multi-objective optimization

Unlike traditional optimization problems where the function to be optimized is unique, multi-objective problems are characterized by the need to optimize several objectives simultaneously [46]. It is even possible that the objectives being considered conflict with each other, where it is not possible to improve one objective without making the other worse. Because of the interplay between the different objectives, in the general multi-objective case, the resolution of the problem implies finding not just one but a set of solutions that represent different tradeoff levels among the criteria to optimize due to the unlikely case of one solution that optimizes all the objectives at the same time.

Formally, we can define the multi-objective optimization problem as:

$$\min_{x} \quad F(x) = (f_1(x), f_2(x), ..., f_n(x)) \tag{2.28}$$

$$\text{s.t.} \quad x \in X \tag{2.29}$$

Where X represents the set of feasible solutions. Without loss of generality, in the following we consider the minimization of the objectives under consideration.

2.3.1 Pareto optimality

In order to get solutions that comprise good tradeoffs from the point of view of the different objectives being considered, it is necessary to define what constitutes a good compromise. We can define this concept in terms of preferences among the solutions with respect to the objective functions. For example, we can take two feasible solutions whose functional values coincide in every objective but one. In this case, it is possible to affirm that the solution whose value in the objective function that differs is lower is preferable to the other, since it improves the alternative solution in at least one objective value and is equal in all remaining objectives. The problem, then, is to find this set of preferable solutions that are not strictly improved by any other solution. This is known in the literature as *Pareto optimal*, and can be defined using the concept of *Pareto Dominance*. Formally, x_1 dominates x_2 if:

$$f_i(x_1) \leq f_i(x_2), \ \forall \ i \in I \tag{2.30}$$

$$\exists i \mid f_i(x_1) < f_i(x_2) \tag{2.31}$$

Pareto Dominance is a strict partial order relation defined over the feasible solutions of a multi-objective problem, and it helps to establish preferences among different solutions, as seen in Figure 2.1.

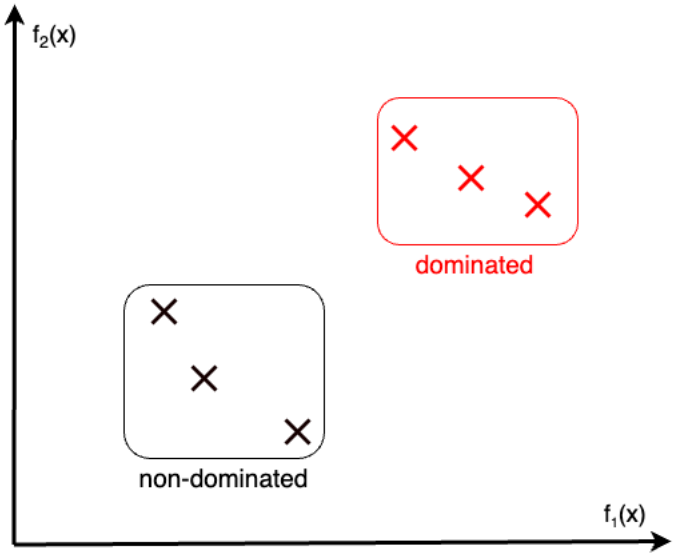

Figure 2.1: Pareto Dominance in a two objective minimization problem

We say then that a solution x^* is Pareto optimal if there exists no other solution $x^{'}$ that dominates x^*.

2.3.2 Resolution methods

Multi-objective optimization problems are hard to solve since calculating an entire Pareto front is often intractable (and of little practical value). Moreover, the challenges in multi-objective problems in the more general case include those of single-objective problems. Therefore, approximate methods, usually based on heuristics or metaheuristics approaches, are usually devised to find an approximate front [28, 75]. The goal in these cases is to determine a set of non-dominated solutions that present interesting levels of tradeoff in the context of the problem under consideration while also considering the stakeholders' preferences.

Two important concepts in multi-objective algorithms are those of *proximity* and *diversity*. These are properties of the approximated Pareto fronts found by heuristic or metaheuristic approaches during their execution [46].

Proximity refers to how close the solutions encountered are with respect to the solutions that conform to the actual Pareto front of the instance under consideration. The approximate Pareto front should be as close to the real Pareto front as possible since this guarantees reasonable quality solutions for a given tradeoff level.

The diversity of an approximate Pareto front increases as more non-dominated solutions with diverse grades of tradeoff across the objectives are found. Diversity is useful when exploring solutions to a multi-objective problem as varying degrees of tradeoff might be worth considering, depending on the specific problem at hand.

In Figure 2.2, both concepts are shown in the context of a minimization problem with two objectives.

The approximation of the Pareto front can be approached in two ways: one is avoiding

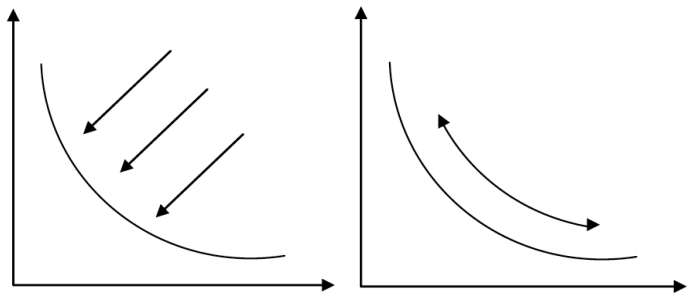

Figure 2.2: Proximity - Diversity (extracted from [57])

calculating the whole front but to ensure that the solutions obtained are Pareto optimal, and the other is to calculate a set of non-dominated solutions without any guarantees of Pareto optimality. Diversity is a goal in both cases, while in the latter proximity is also desired.

In the first case, and when an explicit formulation is available, a popular exact method to calculate the Pareto front is that of epsilon constraint [27]. In the second case, heuristics or metaheuristics approaches are usually devised to seek for solutions with good diversity and proximity [28, 46].

A third alternative is to take a weighted approach [28], which enables the conversion of the problem into a single-objective optimization problem suitable for being solved by exact methods, but it is often difficult to calculate the appropriate weights that will give a good set of solutions with regards to all the objectives under consideration. Moreover, the weighted approach is only valid for convex problems, so it cannot be applied to formulations with discrete variables since, in general, there might exist regions of the front that do not fit any linear combination of weights [27].

2.4 Public transport assignment models

An assignment model represents user behavior: how users satisfy their travel needs using the existing public transportation lines. Users of the system must choose one or more lines from a set of candidate lines that can bring them to their intended destinations. The assignment model is a critical component of any frequency optimization model since user satisfaction is of great importance when measuring the system's performance.

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. These hypotheses affect the different characteristics of the existing assignment models in the literature. The way the users behave directly influences the calculation of measures such as the waiting time and the occupancy of the buses that end users experience. Generally, an assignment model should apply some hypothesis that solves the following issues:

- The lines or combination of lines a user will use for traveling from origin to destination.

- The information users take into account when making such a decision.
- Whether all users with the same travel demands will use the same lines.
- The way users behave when congestion is present in their preferred lines.
- Whether the perception of the travel time is uniform for all users of the same line.

Different assignment models will provide different answers to each of the aforementioned issues, using hypotheses appropriate to the actual case under study.

2.4.1 Classification criteria

Assignment models can be classified according to different criteria [37]:

- *Schedule-based versus frequency-based*: whether users are aware of specific timetables or perceive the lines in terms of headways between subsequent buses departing from a stop. In contexts where the service is so irregular, frequent, or when the timetable information is undisclosed, passengers might not time their arrival at a bus stop. Schedule-based models determine passenger loads on every single line bus, while frequency-based models calculate average loads on the lines under operation.
- *Strategies and hyperpaths*: the strategies the users consider to reach their intended destinations. A travel strategy might include intermediate nodes where users can adapt and change their strategy, i.e., while waiting at a bus stop and boarding the first of a subset of attractive lines. One approach to consider complex strategies is to introduce the concept of *hyperpath* to represent different trajectories from origin to destination instead of a single path on the graph, as is the case when solving the shortest path problem. The strategy of models that allow users to select just one line to reach their destinations (thus deviating from the hyperpath concept) is termed an "all or nothing" assignment. In these cases, the passenger selects a prior single line for which she will wait at the origin bus stop [26].
- *Fixed versus elastic demand*: whether the origin-destination demand is considered fixed and thus independent (and captive) of the services being offered, or elastic, and therefore able to change and adapt in accordance to said services (i.e., by preferring to leverage private transportation such as cars, bikes, or even change the destination due to poor service conditions). Elastic demand introduces some challenges in assignment models since it usually involves a feedback loop to adapt the demand accordingly during the optimization process. Due to this difficulty, fixed-demand models are often chosen in the literature.
- *Path-based versus arc-based.* In path-based assignment models, the lines are explicitly enumerated, either identified in advance or generated during the assignment process, whereas in arc-based models, the arcs are introduced in a sequential manner, and the lines are implicitly enumerated. Path-based approaches allow for more sophisticated models and enable using costs independent from the underlying arc costs, i.e., applying special fares on some of the paths. Still, path computation can

quickly become prohibitive due to the number of alternatives that might need to be enumerated. Arc-based models are less computationally expensive and, when considering strategies, are almost mandatory; thus, they are often the first choice for implementing commercial software [37].

- *Deterministic versus stochastic line choice.* Deterministic models assume that users have the same preferences when choosing the bus to board. Moreover, they also assume that users have perfect information (i.e., travel speed, waiting times) and can manage this information in their best interest. Stochastic models usually employ random utility theory, but the model is generally unable to evaluate exactly these utilities for each user due to factors such as heterogeneity of preferences, incomplete information, or subjective errors in the user's decision-making. A stochastic model would represent the travel time of the users with a random variable, in which case the perception of travel times could be different for users of the same line [61]. More often, the travel time is considered fixed along a given trajectory, and thus, deterministic models are currently the method of choice on most commercial software [37].

- *Simulation-based versus analytical models.* Erratic events, such as the number of passengers waiting at a stop, the actual arrival and departure time of the various buses, and the actual time spent traveling, can affect the daily operation of a transit network in ways that might be better described as random (unpredictable) outcomes. Simulation tools can aid with these challenges, simulating the development of a transit network on any given day. Analytical formulations, on the other hand, provide results as expected values of the output variables of a given assignment and might lack the flexibility of simulation approaches. However, simulation approaches might require several runs to get average values, which can be important when considering different design scenarios. One way analytical formulations can cope with significant random events is by using a rolling horizon approach, i.e., where the analytical model is restarted after some time to forecast the assignment for the next n hours, using the state from the previous run as the input.

Additional classification criteria based on the eventual existence of congestion events over the network are briefly discussed in Section 2.5.

The assignment model embedded in the formulation proposed in this work is the one published in [72], called *optimal strategies*. According to the criteria above, it is a deterministic, frequency-based, arc-based analytical model that represents user behavior by means of a strategy approach with hyperpaths and operates under the assumption of a fixed demand. The *optimal strategies* model is discussed in more detail in Section 3.3.

2.5 Assignment under congestion

If congestion scenarios are considered (due to bus capacities, crowding, and passenger boarding/alighting), then the travel time for a given passenger depends on how other users use the available lines and can not be considered static nor isolated from the rest of the system [15, 50]. The most common approach in these cases is leveraging the user

equilibrium concept. User equilibrium is achieved when no user of a transit network finds it convenient to change its strategy (i.e., hyperpath) anymore to reach its destination. This is usually accomplished by including an equilibrium model embedded into the assignment model. Some assignment models that follow this approach are [22], where the formulation is solved by heuristic means, and [15], where a more formal treatment is presented. In [22], the authors prove that effective frequencies (that is, the actual perceived frequencies by the users of a running system, as opposed to the originally programmed frequencies of the lines, which might not be identical due to service irregularities) can be determined by assigning the demand in the ascending order of the nominal frequencies for each line, concluding that the order in which lines are included in the attractive set can be determined from the start, as it only depends on the travel speeds and is not affected by congestion. Using this result, they propose two algorithms: an approximate one by the linearization of some constraints and a nonlinear algorithm without optimality guarantees. The main drawback of including the models above ([15, 22]) as assignment sub-models in the context of frequency optimization is that the resulting model is complex (becoming an MPEC, Mathematical Program with Equilibrium Constraints) [18].

In [35], another formal approach is proposed to model congestion effects on both in-travel and waiting times. By representing transit stops as complex queueing systems, where passengers arrive randomly with an average rate, an equilibrium model based on queueing is stated. Still, it has proved to be very hard to solve.

An alternative approach to modeling congestion effects formally is to assume that all transit lines can accommodate the demand for their services, effectively ignoring congestion over the system. It is then possible to model the travel times as constants that are not influenced by the network flows, which lessens the complexity of the model. Examples of models of frequency optimization that ensure enough capacity are [19, 20], where the problem is solved in a heuristic manner that is driven by the mathematical properties of an explicit formulation, and [71], where a discomfort penalty affects in-vehicle travel times. Notice that the assumption of meeting all of the demand can be limiting in some contexts where the transit networks operate with services with insufficient, or at the limits of their capacity. Therefore, a more complex approach to congestion, with an explicit modelling of user equilibrium and travel times dependant on the user flow across the different lines, might be desirable under such scenarios.

In the context of this thesis, we propose a model of frequency optimization with strict capacity constraints, in contrast to models that incorporate congestion in the perceived travel times [17]. When an assignment model calculates effective frequencies (which must be higher or equal to the normative ones), in some way, it allows feasible user flows for any frequency values. If the effective frequencies are not capped, then it is always possible to find a feasible solution, even if it is at the cost of prohibitively high waiting times. In some sense, then, this type of model allows a flow assignment that exceeds the capacity. In models such as the one proposed in this thesis, where the capacity constraints are strict, this does not happen: if the capacity limit is reached, then the flow is redirected to other lines, eventually saturating the whole transportation system and thus resulting in an unfeasibility.

2.5.1 Classification criteria

Regarding the treatment of congestion effects, assignment models can be classified according to the following criteria [37]:

- *Uncongested assignment versus user equilibrium*: whether the model considers congestion phenomena to be relevant. As discussed previously, if congestion is considered, the travel times are no longer independent or isolated from the user flow in the rest of the system.
- *Static versus dynamic assignment.* Static models work under the assumption of stable constant flows and quality of service (total travel times, eventually including waiting times) during the assignment period, at least for a sufficiently large period of time where the network does not operate in saturated conditions. This means that each passenger is guaranteed to be able to board the next-arriving vehicle. Dynamic assignment, on the other hand, explicitly deals with capacity restrictions, and the cost of a user's strategy needs to be evaluated at the instant when said choice is committed (i.e., when a passenger boards one of the buses in its strategy).

The assignment model used in this work, *optimal strategies*, does not deal with congestion. Thus, according to the criteria above, it constitutes a model of uncongested static assignment that implicitly assumes user equilibrium.

Chapter 3

Mathematical programming formulation

This chapter describes the proposed formulation for the frequency optimization problem with strict capacity constraints. First, in Section 3.1, the problem is formally defined, as well as the main characteristics and hypotheses considered in the context of this thesis. Then, we introduce the main concepts and modeling approach in Section 3.2.

In order to model user behavior, our formulation incorporates an explicit assignment model [72], which is widely used and accepted in the literature. We present this model in Section 3.3.

Our mathematical formulation is detailed in Section 3.4. We base our formulation on the one proposed in [55]. The objective function represents the interest of the users, who seek to minimize their total travel time, while the operators' interest is expressed through a constraint that sets an upper limit on the total fleet size.

In Subsection 3.4.1, we propose an extension of the model by adding a bus capacity constraint. This leads us, in Subsection 3.4.2, to consider a bilevel formulation able to capture the impact that constraints, such as the bus capacity, have on the nature of the problem.

3.1 Frequency optimization

Since there are many different alternatives to consider when defining the problem of frequency optimization, in this section we describe the aspects deemed relevant and thus included as part of the scope of this thesis.

The main aspects considered are:

- Inclusion of both the interest of the users and the operators.
- Realistic modeling of the behavior of the users when using the provided services.
- The influence of the waiting time on how users behave and perceive the performance of the transportation system.
- Constrain bus capacities.

The users' interest is usually represented by the minimization of travel time. Some works disregard the waiting time of the users at the bus stops, as was stated in Section 1.1, but we do consider it in this thesis.

Operators are generally assumed to be interested in minimizing the overall cost of maintaining and running the resources to support the transportation system. The total fleet size is usually considered a proxy of the cost operators must incur to provide transportation services. They aim to operate with the minimum number of buses while providing a decent level of performance from the users' point of view. Frequencies (as well as the itinerary of the different lines) play a significant role here since, all other things fixed, increasing the frequency of a given line will necessarily require more buses to serve that line and, in turn, will require a bigger crew to operate the buses.

Since there are at least two (potentially conflicting) objectives, some works in the literature have treated the problem as a multiobjective one [2, 38, 58, 65]. Some works [65] have proposed models that represent the multi-objective nature of the problem by using a weighted approach [28]. In such cases, the challenge is to devise an appropriate scale of the objective coefficients, which, for the case of frequency optimization, usually includes a term to represent the interest of the users and another to represent that of the operators.

In this thesis context, the problem's multiobjective nature is not explicitly explored. While taking a weighted approach is relatively straightforward using the proposed formulation (besides scaling considerations), this does not guarantee to achieve a good set of solutions with regard to the different objectives.

The assignment model is the component responsible for modeling the behavior of the users. In the context of this work, we leverage a model widely used in the literature, that seeks to minimize the total travel time of the users (including the waiting time at the bus stops), and allows the use of multiple lines for users of the same OD pair (more on this in Section 3.3).

Finally, one of the main objectives of this thesis is the inclusion of bus capacities in the resulting formulation. A bus capacity constraint involves setting an upper limit to the simultaneous flow that can travel through each line and must be enforced at every segment of the line itinerary. Since the assignment model is responsible for implementing user behavior, the flows of passengers through the lines are part of the final output of the model.

3.2 Graph representation

The most commonly used representation in frequency optimization is by means of graphs. Graphs are mathematical structures well-suited to represent elements such as the itinerary of lines, the street and bus stop network, and even geographical zones and waiting arcs.

Many different graph-based models are proposed in the literature, each according to the designer's specific constraints and objectives. Furthermore, in the case of frequency optimization, a combination of different graph models is required since the same model must represent both the structure of the lines that operate over the provided infrastructure as well as the behavior of the users in terms of line utilization and walking paths [57].

The particular graph model used in this work is detailed in the following sections.

We describe the representation corresponding to the infrastructure (Subsection 3.2.1), the demand (Subsection 3.2.2), and the lines (Subsection 3.2.3) of the transport system.

3.2.1 Infrastructure

The infrastructure consists of the existing bus stops and street segments.

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, so that $N = N^P \cup N^S$. 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. In this manner we define $A = A^T \cup A^B \cup A^L$.

3.2.2 Demand

Without loss of generality, we assume that the demand is generated at the bus stops. Including special nodes that attract or produce demand (commonly called centroid nodes) allows for incorporating zones and walking arcs into the model, but it does not change the underlying nature of the proposed formulation.

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.

3.2.3 Lines

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. We assume that every bus will stop at each bus stop located in the street where the line passes; furthermore, it is out of the scope of this work to model the decision regarding where to locate each bus stop, as well as the design of the underlying street network and infrastructure in general. Figure 3.1 illustrates the graph model.

3.3 Assignment sub-model

The assignment model used in this work is the one proposed in [72], called *optimal strategies*. There are several reasons to represent user behavior using *optimal strategies* in the context of the thesis:

- The model considers the influence of the waiting time at the bus stops on the behavior of the users (i.e., when computing the total travel time). Moreover, the objective is to minimize the total travel time of the users, which is consistent with the hypotheses considered in this work.

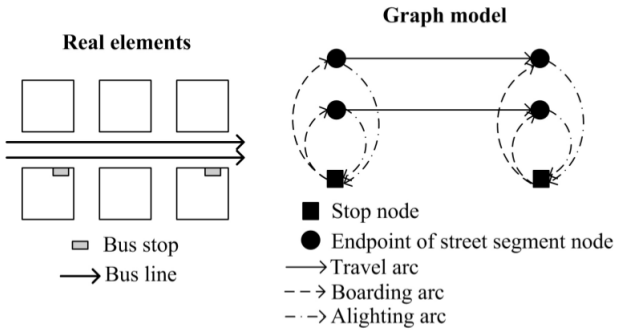

Figure 3.1: Graph model (extracted from [55])

- It is a frequency-based assignment model, which makes it a good choice for embedding it in the proposed frequency optimization model since the frequency information is at the same level of aggregation (as opposed to schedule-based models where the model is more detailed, requiring a timetable of each line as well as more fine-grained demand data [63]).

- It is not an "all or nothing" model [26], permitting the assignment of demand to multiple lines.

- It has an explicit linear and compact mathematical formulation.

- It is a known model in the literature that has been used previously in the context of frequency optimization (sometimes considering line design as well). Solutions from models that do not make the same assumptions cannot be appropriately compared; this thus enables, or at least facilitates, the comparison with solutions reported in different published works.

In *optimal strategies*, a strategy is defined as a set of rules that, when applied, allow users to reach their destinations. In Figure 3.2, an example of a strategy is illustrated.

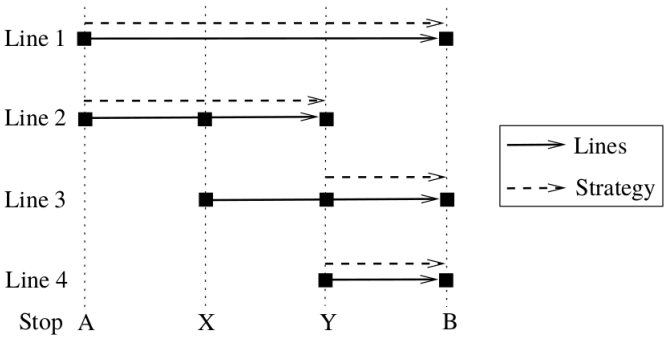

Figure 3.2: Strategy example to travel from A to B (extracted from [57])

Here, a user is originated at bus stop A (or at a centroid close by) and wants to travel to bus stop B (alternatively, to a centroid node close to node B). The strategy is then to wait to board the first line among lines 1 and 2. If line 1 is the first, then the travel requires no transfers, and the user will arrive at his or her desired destination B. If, instead, line 2 is the first, then the user must perform at least one transfer by boarding at bus stop Y, either line 3 or line 4, whichever passes first. Notice that this particular strategy disregards the option of alighting at node X to wait for line 3, something that could be a valid alternative in this example.

The model assumes that a given user selects the strategy that minimizes his total travel time, including the waiting time at the bus stops. To achieve this, it is assumed that users have knowledge of the onboard travel times and frequencies of all the lines of the system, and have the capacity to deal successfully with this information to determine the optimal strategy, which could be quite complex. That information is then used to define 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 shows when considering how the time of a passenger waiting on a stop is calculated, for a set of lines $L = \{l_1, \ldots, l_m\}$ with corresponding frequencies $F = \{f_1, \ldots, f_m\}$. As commonly accepted in the literature [25], the waiting time can then be modeled by a random variable (usually Poisson distributed) of mean value:

$$E(tw) = \frac{\beta}{\sum_{l_i \in L} f_i} \tag{3.1}$$

where β is a parameter that depends on assumptions concerning service regularity. Since the model assumes that passengers take the first bus that arrives to the stop, the probability of using the line l_i, known as the *frequency share rule*, is:

$$P_i = \frac{f_i}{\sum_{l_j \in L} f_j} \tag{3.2}$$

Since the model is probabilistic, a strategy is considered optimal if it minimizes the total expected travel time.

In order to formulate the optimal strategies assignment model, embed the frequency-share rule, and calculate the waiting time, we introduce variables x_a. Then, the expression of the waiting time takes the form

$$\frac{1}{\sum_{a \in A_n^+} f_a x_a} \tag{3.3}$$

where x_a is a binary variable that indicates whether arc a is part of the optimal strategy, A_n^+ is the set of outgoing arcs from node n and f_a is the frequency (number of buses per unit of time) of the line that corresponds to the boarding arc a.

In this context, the frequency share rule is written as

$$v_a = \frac{V_n f_a x_a}{\sum_{a' \in A_n^+} f_{a'} x_{a'}} \tag{3.4}$$

where V_n represents the flow on node n and v_a is the flow through arc $a \in A$. The authors of the original work simplify the model by introducing the following change of variables:

$$w_n = \frac{V_n}{\sum_{a \in A_n^+} f_a x_a} \tag{3.5}$$

and eliminating variable x_a. For further details, we refer interested readers to the original publication [72].

Since in the *optimal strategies* model congestion is not considered, the behavior of a given user is independent from that of any other, which allows to formulate the assignment model, for a single OD pair, in the following way:

$$\min_{v,w} \quad \sum_{a \in A} c_a v_a + \sum_{n \in N^P} w_n \tag{3.6}$$

$$\text{s.t.} \quad \sum_{a \in A_n^+} v_a - \sum_{a \in A_n^-} v_a = b_n \qquad \forall\, n \in N, \tag{3.7}$$

$$v_a \leq f_a w_n \qquad \forall\, n \in N^P, a \in A_n^+, \tag{3.8}$$

$$v_a \geq 0 \qquad \forall\, a \in A \tag{3.9}$$

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 (3.6) states the intention of the users of the system, that is, to minimize their total travel time (sum of onboard travel time and the waiting time at the stops). The flow conservation constraint (3.7) guarantees that all users can reach their destinations. Constraint (3.8) splits the demand at each stop node 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 the optimal strategy, and the constraint verifies with equality, restoring the frequency share rule expression.

Note that the model, as stated above, does not allow for walk arcs between bus stops, so the assumption is that if needed, transfers can only take place at the alighting bus stop.

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 (3.8) 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* [60] 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.

3.4 Frequency optimization model

We base our formulation on the one proposed in [55], which can be stated as follows:

$$\min_{f,v,w} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{3.10}$$

$$\text{s.t.} \quad \sum_{l\in L} f_l \sum_{a\in l} c_a \leq B, \tag{3.11}$$

$$\sum_{a\in A_n^+} v_{ak} - \sum_{a\in A_n^-} v_{ak} = b_{nk} \qquad \forall\, n \in N, k \in K, \tag{3.12}$$

$$v_{ak} \leq f_a w_{nk} \qquad \forall\, a \in A_n^+, n \in N^P, k \in K, \tag{3.13}$$

$$v_{ak} \geq 0 \qquad \forall\, a \in A, k \in K, \tag{3.14}$$

$$f_l \geq 0 \qquad \forall\, l \in L. \tag{3.15}$$

$$w_{nk} \geq 0 \qquad \forall\, n \in N^P, k \in K, \tag{3.16}$$

where B is the maximum fleet size allowed across all lines. In formulation (3.10 - 3.16), the objective function is that of the users, which intend to minimize their total travel times, while the interest of the planners, seeking to minimize operational costs, is considered as a constraint in (3.11). The assignment model is included in constraints (3.12 - 3.14), now expanded to consider each demand pair k.

The nonlinearity in formulation (3.10 - 3.16) arises from the inclusion of the demand split constraint (3.13). Although in the context of the original optimal strategies assignment model (3.6 - 3.9), the demand split constraint is linear since the frequencies f_a are inputs to the model that were already set by the operator of the service, in the broader context of frequency optimization, the constraint is nonlinear since both terms f_a and w_{nk} are decision variables that need to be determined.

Another shortcoming of the previous formulation is the lack of an upper limit to the frequencies of the different lines, which is unrealistic for technical reasons. A lower limit constraint for the frequencies (that is, a lower limit higher than 0) could also be desirable to ensure a certain minimum level of service to the users of the system; alternatively, this could also be modeled by having an upper limit constraint on the waiting times.

The frequency optimization model proposed in [55] solves the aforementioned issues. It is stated as follows:

$$\min_{y,v,w} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{3.17}$$

$$\text{s.t.} \quad \sum_{l\in L}\sum_{f\in\Theta} \theta_f y_{lf} \sum_{a\in l} c_a \leq B, \tag{3.18}$$

$$\sum_{f\in\Theta} y_{lf} = 1 \qquad \forall\, l \in L, \tag{3.19}$$

$$\sum_{a\in A_n^+} v_{ak} - \sum_{a\in A_n^-} v_{ak} = b_{nk} \qquad \forall\, n \in N, k \in K, \tag{3.20}$$

$$v_{ak} \leq \theta_{f(a)} w_{nk} \qquad \forall\, a \in A_n^+, n \in N^P, k \in K, \tag{3.21}$$

$$v_{ak} \geq 0 \qquad \forall\, a \in A, k \in K, \tag{3.22}$$

$$v_{ak} \leq \delta_k y_{l(a)f(a)} \qquad \forall\, a \in A^B, k \in K, \tag{3.23}$$

$$y_{lf} \in \{0,1\} \qquad \forall\, l \in L, f \in \Theta. \tag{3.24}$$

Formulation (3.17 - 3.24) is a linear transformation of formulation (3.10 - 3.16), with 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. 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 from that node as possible values of Θ. In that way, some loss of precision is introduced, but the authors state that in real systems, it is usually convenient to consider a reduced set of frequency values due to service coordination and fleet management issues. Figure 3.3 illustrates the changes introduced in the graph model by using a discretized domain of frequencies.

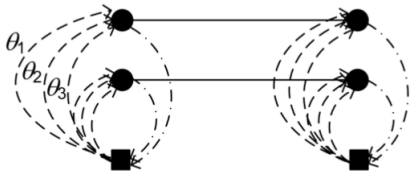

Figure 3.3: Discretized domain of frequencies (extracted from [55])

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 indicate the line frequencies some notation is needed: $f(a)$ 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.

With regards to the original formulation (3.10 - 3.16) a new constraint (3.19) enforces the fact that each line must have exactly one frequency associated, while constraint (3.23) 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. We also note that some of the original expressions are now written in terms of Θ whenever an actual frequency value is required.

This results in a mixed integer linear formulation where the primary source of complexity is the existence of binary variables and the increase in the size of the underlying graph model due to the addition of new boarding arcs (one per possible frequency value), which stems from the discretization of the domain of frequencies. Nevertheless, in [55], an instance corresponding to an actual city, comprising 84 nodes, 143 edges, 378 OD pairs, and 13 lines, was solved exactly with the proposed formulation.

3.4.1 Adding the bus capacity constraint

The assignment sub-model embedded in formulation (3.17 - 3.24) assumes that there is sufficient capacity to carry all the passengers that desire to use any line. There is no additional constraint in the formulation that considers the capacity of the lines, something unrealistic in systems that exhibit a 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, line flows in principle could be imposed by adding the following constraint:

$$\sum_{k \in K} v_{ak} \leq \sum_{f \in 1..m} y_{l(a)f} \theta_f \omega \quad \forall a \in A^T \tag{3.25}$$

where, as previously stated, $y_{l(a)f}$ is a binary variable specifying whether the line of arc a has frequency f, and θ_f represents the actual value of said frequency f.

However, adding this constraint directly in formulation (3.17 - 3.24) 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 (3.25) is active for their corresponding travel arcs. This represents the optimal strategy.
- Other alternative hyperpaths, whose cost according to expression (3.6) 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 that is saturated. These alternative non-optimal hyperpaths might, for example, represent cases where a user, instead of boarding the bus that would take her faster to her destination, decides to let it pass in order to improve (minimize) the total travel time of other users of the system that would contribute more heavily on the total travel time of the whole system objective (i.e. because the other users need to travel longer distances than herself).

This leads us to the concepts of *line planning with route assignment* (LPRA) and *line planning with route choice* (LPRC), first defined in [40]. 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. On the other hand, assignment models such as the one used in this work imply an LPRC approach, where each user chooses the route that best fits his or her own expectations. Adding constraint

(3.25) directly into the formulation would violate the LPRC approach, as users would need to consider a priori lines that must conform with the new bus capacity constraint (planners concern) rather than choosing the lines in an egoistic way. In general, adding any constraint that may impact the variables that model user behavior and that are not required by the hypothesis of the assignment model would defeat the purpose of the model, since users would behave in such a way as to pursue the optimization of some global optimum that benefits the formulation in place but not necessarily their own interests (objectives).

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

- Assuming that the planner ensures sufficient capacity on the lines that the users want to use, which is done by setting appropriate frequencies on the corresponding lines.

- Modeling a congested system through an assignment sub-model that 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 to wait for a different line.

There have been very few works studying the problem of frequency optimization with assignment under congestion [17, 62], probably because this second alternative needs to consider an equilibrium assignment sub-model [15, 22, 49] embedded into the frequency optimization, which is considerably more complex than the first approach [33]. Furthermore, to the best of our knowledge, there is no formal criterion for deciding between both approaches from a modeling point of view. In practice, constraints related to the capacity of the infrastructure, budget, and policy come into play to determine whether it is possible to operate a non-congested system.

In this work, we follow the first approach. Ensuring sufficient capacity is possible by increasing the frequencies of the saturated lines. This is a decision of the planners (variable y in expression (3.25)) who should take into account the decisions of the users (variables v and w). This leads us to consider a bilevel optimization model, where the upper level represents the decisions of the planner, and the lower level represents the decisions of the users.

3.4.2 Bilevel mathematical programming formulation

If constraint (3.25) is added to formulation (3.17 - 3.24), it 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. These kinds of scenarios are usually modeled using bilevel mathematical programs [7, 18, 31].

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

$$\min_{y,v,w} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{3.26}$$

$$\text{s.t.} \quad \sum_{l\in L}\sum_{f\in\Theta} \theta_f y_{lf} \sum_{a\in l} c_a \leq B, \tag{3.27}$$

$$\sum_{f\in\Theta} y_{lf} = 1 \qquad \forall\ l \in L, \tag{3.28}$$

$$\sum_{k\in K} v_{ak} \leq \sum_{f\in\Theta} y_{l(a)f}\theta_f \omega \qquad \forall\ a \in A^T, \tag{3.29}$$

$$y_{lf} \in \{0,1\} \qquad \forall\ l \in L, f \in \Theta, \tag{3.30}$$

$$\min_{v,w} \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{3.31}$$

$$\text{s.t.} \sum_{a\in A_n^+} v_{ak} - \sum_{a\in A_n^-} v_{ak} = b_{nk} \qquad \forall\ n \in N, k \in K, \tag{3.32}$$

$$v_{ak} \leq \theta_{f(a)} w_{nk} \qquad \forall\ a \in A_n^+, n \in N^P, k \in K, \tag{3.33}$$

$$v_{ak} \leq \delta_k y_{l(a)f(a)} \qquad \forall\ a \in A^B, k \in K, \tag{3.34}$$

$$v_{ak} \geq 0 \qquad \forall\ a \in A, k \in K. \tag{3.35}$$

where the upper level (3.26 - 3.30) represents decisions of the planners while the lower level (3.31 - 3.35) represents decisions of the users, that is, the assignment sub-model taking as input the fixed frequencies $\theta_{f(a)}$. The objective function of both levels is the same and only considers the interest of the users, which is to minimize the overall travel time. Arguably, the fleet size constraint (3.27) could be modeled as another objective to minimize at the upper level, which would lead us to consider a multi-objective bilevel formulation, increasing the complexity of the formulation [38].

The planners can ensure sufficient capacity on the lines that users want to use by adjusting the frequencies according to constraint (3.29). In this manner, users are assumed to perceive unlimited capacities on the lines they might take. This implies that bus capacities are not considered in the assignment sub-model, but rather, the bus capacity is included as a constraint of the frequency optimization model. This is not always feasible since there might be scenarios where the demand for some of the lines is too high and cannot be fulfilled due to technical limitations of the transportation infrastructure. In such cases, modeling a congested system by adding the bus capacity constraint in the assignment sub-model, as was discussed in Subsection 3.4.1, might be necessary, even though the complexity of the resulting formulation would be higher. To the best of our knowledge, no works published in the literature provide an analysis comparing the performance of two alternative transportation models that only differ in how bus capacities are handled.

Formulation (3.26 - 3.35) is classified as Discrete Continuous Linear Bilevel (DCLB) [74] 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, could be solved to optimality. Some of the more popular reformulation strategies for achieving this are:

- Using the Karush-Kuhn-Tucker conditions to substitute the lower level problem and therefore removing the distinction among the different levels. Due to the complementarity term, which 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. This approach has been described and used in [7, 31].

- 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 [5, 9, 34].

In the present work, formulation (3.26 - 3.35) was transformed into a single level formulation using the first approach, that is, by replacing the lower level with 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. The second approach, while mathematically equivalent to using the KKT conditions [5], does not blend itself to an easy linearization, due to the introduction of the strong duality equality as a nonlinear constraint.

By replacing the lower level with its optimality conditions, variables that represent the decisions of the users (v and w) are restricted to take values that solve problem (3.31 - 3.35). Therefore, the whole model will adjust the frequency values (variables y) so as to respect the constraints that are directly included in the upper level (among them, bus capacity) as well as the optimality conditions that represent the (uncapacitated) lower level problem. By applying concepts of duality theory [10] to the optimal strategies assignment model (adapted to our formulation), we calculate the dual of problem (3.31 - 3.35) as follows:

$$\max_{\pi,\mu,\nu} \quad \sum_{k\in K}\sum_{n\in N^P} b_{nk}\pi_{nk} - \sum_{k\in K}\sum_{a\in A_n^{B+}} \delta_k y_{l(a)f(a)}\mu_{ak} \tag{3.36}$$

$$\text{s.t.} \quad \pi_{ik} - \pi_{jk} \le c_a \qquad \forall\ a=(i,j)\in A - A^B, k\in K, \tag{3.37}$$

$$\pi_{ik} - \pi_{jk} - \mu_{ak} - \nu_{ak} \le c_a \qquad \forall\ a=(i,j)\in A^B, k\in K, \tag{3.38}$$

$$\sum_{a\in A_n^{B+}} \theta_{f(a)}\nu_{ak} \le 1 \qquad \forall\ n\in N, k\in K, \tag{3.39}$$

$$\mu_{ak}, \nu_{ak} \ge 0 \qquad \forall\ a\in A^B, k\in K \tag{3.40}$$

where π_{nk}, ν_{ak}, and μ_{ak} are the dual variables corresponding to constraints (3.32), (3.33), and (3.34), respectively.

The complementary slackness conditions are

$$s^1_{ak}\nu_{nk} = 0 \qquad \forall\ a \in A, n \in N, k \in K, \tag{3.41}$$
$$s^2_{ak}\mu_{ak} = 0 \qquad \forall\ a \in A, k \in K, \tag{3.42}$$
$$t^1_{ak}\nu_{ak} = 0 \qquad \forall\ a \in A - A^B, k \in K, \tag{3.43}$$
$$t^2_{ak}\nu_{ak} = 0 \qquad \forall\ a \in A^B, k \in K, \tag{3.44}$$
$$t^3_{nk}w_{nk} = 0 \qquad \forall\ n \in N, k \in K \tag{3.45}$$

where s^1_{ak} and s^2_{ak} are slack variables associated to inequality constraints (3.33) and (3.34), respectively, and t^1_{ak}, t^2_{ak} and t^3_{nk} are slack variables associated to the inequality constraints (3.37), (3.38) and (3.39), respectively.

The complementary slackness conditions can be linearized by applying the big-M method [31], which uses the disjunctive nature of the conditions and proposes to substitute each product xy by equations

$$x \leq Mz, \tag{3.46}$$
$$y \leq (1 - z)M \tag{3.47}$$

where z is a binary variable, and M is a sufficiently high positive value.

Using the KKT conditions, the resulting MILP model, equivalent to (3.26 - 3.35), is obtained by substituting the lower level (3.31 - 3.35) with its original constraints (3.57 - 3.60), the constraints of its dual (3.36 - 3.40), and the linearized version of the complementary slackness conditions (3.61 - 3.70). The formulation is shown below:

$$\min_{y,v,w,\pi,\mu,\nu,s,t} \quad \sum_{k\in K}(\sum_{a\in A} c_a v_{ak} + \sum_{n\in N^P} w_{nk}) \tag{3.48}$$
$$\text{s.t.} \quad \sum_{l\in L}\sum_{f\in\Theta} \theta_f y_{lf} \sum_{a\in l} c_a \leq B, \tag{3.49}$$
$$\sum_{f\in\Theta} y_{lf} = 1 \qquad \forall\ l \in L, \tag{3.50}$$
$$\sum_{k\in K} v_{ak} \leq \sum_{f\in\Theta} y_{l(a)f}\theta_f\omega \qquad \forall\ a \in A^T, \tag{3.51}$$
$$\sum_{a\in A^+_n} v_{ak} - \sum_{a\in A^-_n} v_{ak} = b_{nk} \qquad \forall\ n \in N, k \in K, \tag{3.52}$$
$$v_{ak} \leq \theta_{f(a)} w_{nk} \qquad \forall\ a \in A^+_n, n \in N^P, k \in K, \tag{3.53}$$
$$v_{ak} \geq 0 \qquad \forall\ a \in A, k \in K, \tag{3.54}$$
$$v_{ak} \leq \delta_k y_{l(a)f(a)} \qquad \forall\ a \in A^B, k \in K, \tag{3.55}$$
$$y_{lf} \in \{0, 1\} \qquad \forall\ l \in L, f \in \Theta, \tag{3.56}$$
$$\pi_{ik} - \pi_{jk} \leq c_a \qquad \forall\ a = (i, j) \in A - A^B, k \in K, \tag{3.57}$$
$$\pi_{ik} - \pi_{jk} - \mu_{ak} - \nu_{ak} \leq c_a \qquad \forall\ a = (i, j) \in A^B, k \in K, \tag{3.58}$$

$$\sum_{a \in A_n^{B+}} \theta_{f(a)} \nu_{ak} \leq 1 \qquad \forall\, n \in N, k \in K, \quad (3.59)$$

$$\mu_{ak}, \nu_{ak} \geq 0 \qquad \forall\, a \in A^B, k \in K, \quad (3.60)$$

$$\theta_{f(a)} w_{ik} - v_{ak} \leq s^1_{ak} M \qquad \forall\, a = (i,j) \in A^B, k \in K, \quad (3.61)$$

$$\nu_{ak} \leq (1 - s^1_{ak}) M \qquad \forall\, a \in A^B, k \in K, \quad (3.62)$$

$$\delta_k y_{l(a)f(a)} - v_{ak} \leq s^2_{ak} M \qquad \forall\, a \in A^B, k \in K, \quad (3.63)$$

$$\mu_{ak} \leq (1 - s^2_{ak}) M \qquad \forall\, a \in A^B, k \in K, \quad (3.64)$$

$$c_a - \pi_{ik} + \pi_{jk} \leq t^1_{ak} M \qquad \forall\, a = (i,j) \in A - A^B, k \in K, \quad (3.65)$$

$$v_{ak} \leq (1 - t^1_{ak}) M \qquad \forall\, a \in A - A^B, k \in K, \quad (3.66)$$

$$c_a - \pi_{ik} + \pi_{jk} + \mu_{ak} + \nu_{ak} \leq t^2_{ak} M \qquad \forall\, a = (i,j) \in A^B, k \in K, \quad (3.67)$$

$$v_{ak} \leq (1 - t^2_{ak}) M \qquad \forall\, a \in A^B, k \in K, \quad (3.68)$$

$$1 - \sum_{a \in A_n^{B+}} \theta_{f(a)} \nu_{ak} \leq t^3_{nk} M \qquad \forall\, n \in N, k \in K, \quad (3.69)$$

$$w_{nk} \leq (1 - t^3_{nk}) M \qquad \forall\, n \in N, k \in K, \quad (3.70)$$

$$s^1_{ak} \in \{0,1\} \qquad \forall\, a \in A, k \in K, \quad (3.71)$$

$$s^2_{ak} \in \{0,1\} \qquad \forall\, a \in A^B, k \in K, \quad (3.72)$$

$$t^1_{ak} \in \{0,1\} \qquad \forall\, a \in A - A^B, k \in K, \quad (3.73)$$

$$t^2_{ak} \in \{0,1\} \qquad \forall\, a \in A^B, k \in K, \quad (3.74)$$

$$t^3_{nk} \in \{0,1\} \qquad \forall\, n \in N, k \in K \quad (3.75)$$

Despite being a single level MILP, formulation (3.48 - 3.75) is not necessarily easy to solve. As was mentioned in Section 2.2, bilevel problems are intrinsically hard. Note that the resulting formulation comprises a high number of binary variables required to linearize the nonlinear expressions.

Chapter 4

Numerical experiments

In this chapter, we perform several numerical experiments using the proposed formulation as well as other alternatives. The main goals and research questions are:

- Study the bilevel nature of the capacitated frequency optimization problem. Is it necessary to employ a bilevel approach when seeking solutions? How does the bilevel nature and the interplay of planners and users affect the solutions encountered?
- How flexible is the proposed formulation when considering the addition of new constraints? What kind of additional constraints could be added, and how would they affect the solutions?
- Study alternative formulations that could aid when exploring other aspects of reality, such as getting measures of the total fleet an operator must provide in order to guarantee a certain level of service, or ensuring that individual users do not experience arbitrarily high waiting times at the bus stops.
- Explore the feasibility of applying the model to a real case, with results published in the literature.
- Up to which size, in terms of the underlying graph, are instances solvable to optimality?
- Is the formulation suitable to be used in a real life context, as part of the process of decision-making in the determination of frequencies?

Instances of different sizes and characteristics were used to explore the aforementioned points.

4.1 Small instance

In order to illustrate the application of the bilevel model explained in Chapter 3, we show in Figure 4.1 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

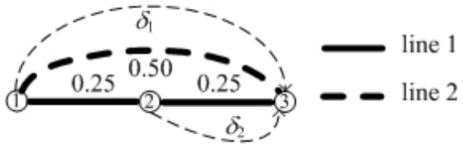

Figure 4.1: Small instance case

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 4.1: 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

Table 4.1 shows the results of applying three different variants of the model denoted by formulation (3.48 - 3.75) to the example of Figure 4.1, where τ (calculated in (3.48)) is the total travel time of the optimal solution and β (calculated in (3.49)) its corresponding fleet size; it also shows the line capacity (as defined in expression (3.25)) and the critical flow of each line (defined as the maximum flow v_a on the arcs of the line). Even though the model has a large number of variables, due to the small size of the instance, the execution times were negligible.

4.1.1 Comparison of uncapacitated and single level capacitated models

The first line of the table shows the results of applying the uncapacitated model (3.17 - 3.24). 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) according to the flow splitting constraint (3.21). Although the critical flow of line l_1 exceeds its capacity, since the model is uncapacitated, this does not turn the solution infeasible but rather illustrates the lack of realism that uncapacitated model solutions might face. This is shown in Figure 4.2.

When we consider bus capacities in the original uncapacitated model (second line of the table, Figure 4.3), by 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}).

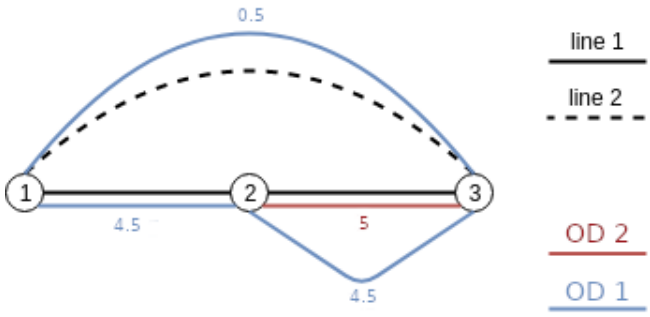

Figure 4.2: Flows in the uncapacitated model

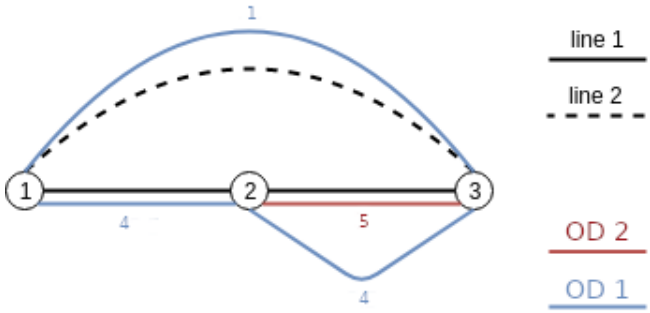

Figure 4.3: Flows in the capacitated single-level model

The examples above show, through a numerical application, the consequences of solving the capacitated problem in a straightforward (non realistic) way. When the bilevel model (3.48 - 3.75) is applied to the same case, the problem is unfeasible. This is due to the fleet size constraint, which 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 set the frequencies in such a way as to redistribute the flows in order to respect the line capacities. This difficulty was already noted in [19]. To overcome this issue, we identify two approaches in the literature:

- Soften the bus capacity constraint, by moving it as a term of the objective function [19].
- Allow the model to increase the fleet size, by including its respective constraint in the objective function [52].

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. In the case of the second approach, the assumption is 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 [58].

4.1.2 Determining the 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 the upper level objective, subject to a constraint of maximum travel time; that is, swapping objective function (3.26) and constraint (3.27).

The results of applying this model to the small instance can be found in the third line of Table 4.4, where we state a maximum travel time equal to 4.8 (the optimal value of the uncapacitated run). 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 that respects the bus capacity constraint while at the same time producing a total travel time that is no worse than the one corresponding to the uncapacitated case, the fleet size should be increased by 15%.

4.1.3 Main findings

In this section, we have shown that models of frequency optimization that disregard bus capacities can easily give misleading results in contexts where passenger capacity is important. This, which was made evident with a minimal case, can become more serious in scenarios representing actual cities, where some lines might end up serving a disproportionate amount of demand that would cause congestion, higher waiting times, and a different assignment of flows over the network, invalidating all service performance metrics that are part of the output of an assignment model.

We have also shown the perils of modeling bus capacities within a single level formulation that includes explicit user behavior. This is also problematic since, in real-life contexts, users cannot be steered by the operators on which lines to board to reach their destinations. Again, this presents a discrepancy with reality, where the assignment of flows would continue to honor users' egoistic interests, invalidating thus most of the performance metrics of the transportation system under consideration.

Finally, even in cases where a feasible solution can not be obtained using a bilevel approach (for example, due to the interplay between bus capacities and the total fleet size), we showed that it's still possible to leverage the formulation to analyze and understand how much to relax some of the original constraints (i.e., the fleet size) in order to provide an acceptable level of service for the users of the system.

4.2 Mandl test case

The test case of Mandl represents part of the public transportation system from an unspecified city in Switzerland [43, 53]. The network is composed of 15 nodes and 21 arcs. The origin-destination matrix is symmetric and very dense, with 76% of non-zero elements. This high-density OD matrix causes the bus capacity constraint to become active under certain line and frequency configurations, a scenario that is of interest in this thesis.

It is one of the most popular test cases in the literature, having been used as a benchmark by several authors [16, 29, 43, 66, 77]. It is also included in the OR Library repository [1]. Despite its widespread use, there is little information about the construction procedure of the test case, and it presents certain characteristics that are not among those expected in a typical real context. It has been used both for validation purposes and as a way to compare the efficiency of different resolution methods.

In Figure 4.4, Mandl's network topology is depicted. The nodes represent bus stops, the arcs represent street (travel) arcs between the stops, and the numbers close to the arcs indicate their corresponding travel times given in minutes.

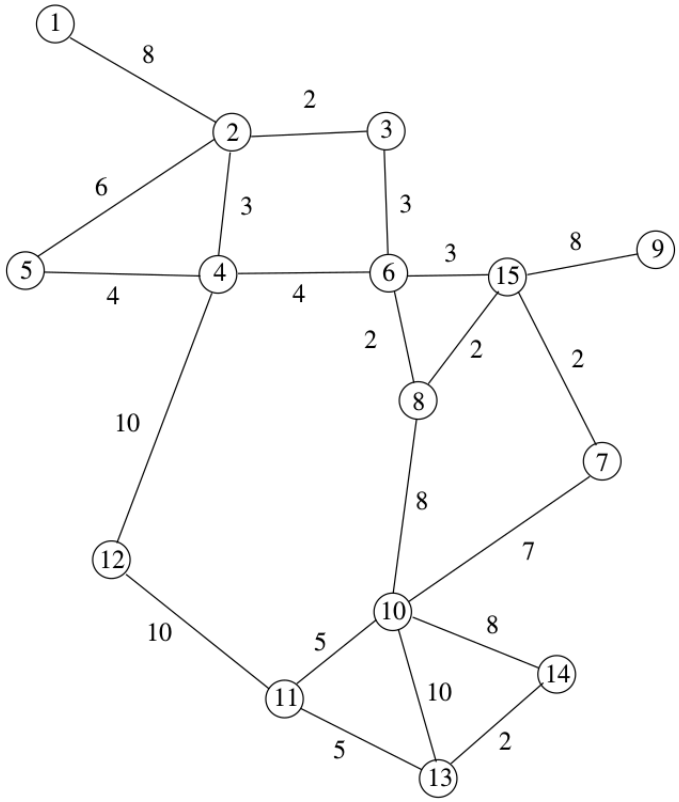

Figure 4.4: Mandl's network

The origin-destination matrix is shown in Figure 4.5, where the entries are expressed in amount of travels per day. We also note that the demand originates and ends at the bus stops.

4.2.1 Prerequisites for the comparison with published results

In this section, we establish the basis for a coherent comparison of results obtained by different resolution methods in the context of frequency optimization. Comparing different algorithms is of great importance in order to explore the strengths and weaknesses of the different approaches in terms of both performance and the quality of the solutions

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	0	400	200	60	80	150	75	75	30	160	30	25	35	0	0
2	400	0	50	120	20	180	90	90	15	130	20	10	10	5	0
3	200	50	0	40	60	180	90	90	15	45	20	10	10	5	0
4	60	120	40	0	50	100	50	50	15	240	40	25	10	5	0
5	80	20	60	50	0	50	25	25	10	120	20	15	5	0	0
6	150	180	180	100	50	0	100	100	30	880	60	15	15	10	0
7	75	90	90	50	25	100	0	50	15	440	35	10	10	5	0
8	75	90	90	50	25	100	50	0	15	440	35	10	10	5	0
9	30	15	15	15	10	30	15	15	0	140	20	5	0	0	0
10	160	130	45	240	120	880	440	440	140	0	600	250	500	200	0
11	30	20	20	40	20	60	35	35	20	600	0	75	95	15	0
12	25	10	10	25	15	15	10	10	5	250	75	0	70	0	0
13	35	10	10	10	5	15	10	10	0	500	95	70	0	45	0
14	0	5	5	5	0	10	5	5	0	200	15	0	45	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4.5: Mandl's origin-destination matrix (extracted from [43])

encountered. When results from other resolution methods are available, they also allow to validate the correctness of a new proposal. In the context of this work, we are mainly interested in validating the proposed formulation, as well as studying the characteristics of the solutions found.

There are several challenges when attempting to compare solutions. The main issues are:

- Lack of standard test cases. Most publications about frequency optimization where results are reported use fictitious or partially fictitious cases. It is hard to assemble a test case with real characteristics, particularly when the objective is to study specific properties of the models under consideration, like congestion.
- Many different variants of published models that solve the frequency optimization problem. The hypotheses tend to differ since they are usually relevant to the specific test cases under consideration. Differences in the constraints or in the treatment of the operator and user objectives prevent, in many cases, the comparison across solutions obtained by applying different models.
- Even when the optimization models are identical with regard to the decision variables, constraints, and objective functions, the behavior of the users could be modeled in different ways, implying the use of different assignment models. If users do not react the same way when the same set of lines is presented, results are no longer comparable: this must be since the assignment sub-model is responsible for calculating important measures such as the flow of users through the different lines, or the waiting times at the bus stops, which in turn impact the total travel time as perceived by the users and capacity constraints, among other considerations.

Another challenge when comparing solutions obtained by an exact method, as in the case of the current work, is the lack of exact resolution methods proposed in the literature,

especially when contrasted with the number of works using heuristic or metaheuristic approaches, where optimality is not guaranteed, and no optimality gap is provided.

In the present work, we use the results published by Baaj and Mahmassani [43] in order to compare the solutions obtained by applying formulation (3.48 - 3.75) to the test case of Mandl. Certain differences between the model and algorithm proposed by Baaj and Mahmassani [43] need to be taken into account when interpreting the results:

- Transfers (when users need to board two or more buses to reach their destination) are considered both in the evaluation of the objective function (by means of a penalty per transfer taken) and in the behavior of the users.

- The assignment model, implemented by the Transit Route Analyst algorithm (TRUST) [42] presents several differences with respect to the optimal strategies assignment model. TRUST's assignment choice considers two criteria: the number of transfers necessary to reach the destination and the total travel time (including waiting times) of the different alternative lines. If there is a tie in the number of transfers (or no transfers are needed), then the second criterion comes into play, and users will seek to choose lines from a set of lines whose travel times are within a particular range of the in-vehicle travel time. Then, for that set of lines, a frequency share rule is applied that resembles the one used in the optimal strategies assignment model (Equation 3.2). Moreover, the assignment sub-model is implemented as a procedural routine instead of as a formal mathematical programming formulation. The TRUST assignment model is explained in more detail in Appendix A.

- The problem considers both the design of the lines (itineraries) and the setting of frequencies. For the frequency setting, no explicit search in the domain of frequencies is performed; instead, frequencies are set at their minimum to get a feasible configuration that respects the maximum load factor of the different lines, that is, the maximum number of standing passengers allowed over the seating capacity of a bus. Furthermore, a heuristic procedure attempts to calculate the minimal frequencies required for a given set of lines, but no proof is presented about the convergence of such a procedure despite the authors claiming empirical evidence.

Despite the differences stated previously, there are enough shared characteristics between the two approaches to make such a comparison feasible. The set of constraints is roughly the same but for a restriction on the maximum number of transfers (two in the case of Baaj and Mahmassani, unbounded in the present work). Bus capacities are also considered in [43], as well as waiting times. The users' objective is represented in the same way, that is, as a minimization of their total travel times, including waiting times. Moreover, the time over the travel arcs is constant, and independent of the flows. The interest of the operators is also expressed in the total fleet number required to operate the services. Regarding user behavior, both assignment models consider the assignment of demand of the same OD pair to multiple lines by means of a split rule based on the line frequencies.

Baaj and Mahmassani's algorithm relies on three subroutines that compute various measures. During the first stage, a greedy algorithm generates an initial set of lines

Table 4.2: Mandl - Line itineraries used in the OPT instances

Line	Itinerary
1	$(1, 2, 3, 6, 8, 10, 11, 13)$
2	$(5, 4, 6, 8, 15, 7)$
3	$(12, 4, 5, 15, 9)$
4	$(13, 14, 10)$

(Route Generation Algorithm). The solutions from this step are evaluated using the TRUST algorithm, which implements the assignment model. A local improvement phase occurs next (Route Improvement Algorithm), and then the solutions are again evaluated using TRUST. The algorithm steps are depicted in Figure 4.6. We highlight in red the phases relevant to frequency assignment.

The algorithm by Baaj and Mahmassani is based on heuristics, so there are no optimality guarantees. This means there could be better solutions than those reported in the publication for the proposed model and assignment hypotheses.

4.2.2 Results comparison

In this section we compare the results obtained from applying formulation (3.48 - 3.75) to the test case of Mandl with those published by Baaj and Mahmassani [43] and those published by Mandl in the original publication [53] using the same test case. The main goal of this section is to validate the proposed approach while also providing some sense of the diversification of the solutions encountered by various runs of the exact method.

The four lines used in Mandl [53] are depicted in Table 4.2. It is also the set of lines considered in all the experiments reported in this section using formulation (3.48 - 3.75).

This set of lines satisfies 100% of the total demand. Unfortunately, in Baaj and Mahmassani [43], the itineraries of the resulting lines and associated frequencies generated by the algorithm are not mentioned. However, the authors state that their solutions also cover 100% of the demand.

In the original work, three runs of Baaj and Mahmassani's algorithm were performed, where each run differed either by the model parameters used or by the line construction strategies employed. In this way, the authors attempted to obtain solutions with different levels of tradeoff with regard to the objectives of users and operators, although the final solutions are rather concentrated on a specific region of the Pareto front [57]. Furthermore, due to the artificial characteristics of the test case, it is not possible to state which part of the Pareto front corresponds to practicable solutions.

In all runs, the bus seating capacity was set as 40, and a bus load factor of 1.25 was chosen. The bus load factor is a coefficient that, when multiplied by the seating capacity, gives the total amount of passengers (both standing and seated) that can travel in a bus, denoted in our model by ω. This configuration then translates to setting $\omega = 50$. A transfer time penalty was imposed, but for the purposes of this section we do not include it in the reported total travel times since our proposed formulation does not penalize transfers. We still show the amount of demand served directly and by one transfer. Notice that while our formulation enables transfers, it does not include a way of extracting that information from

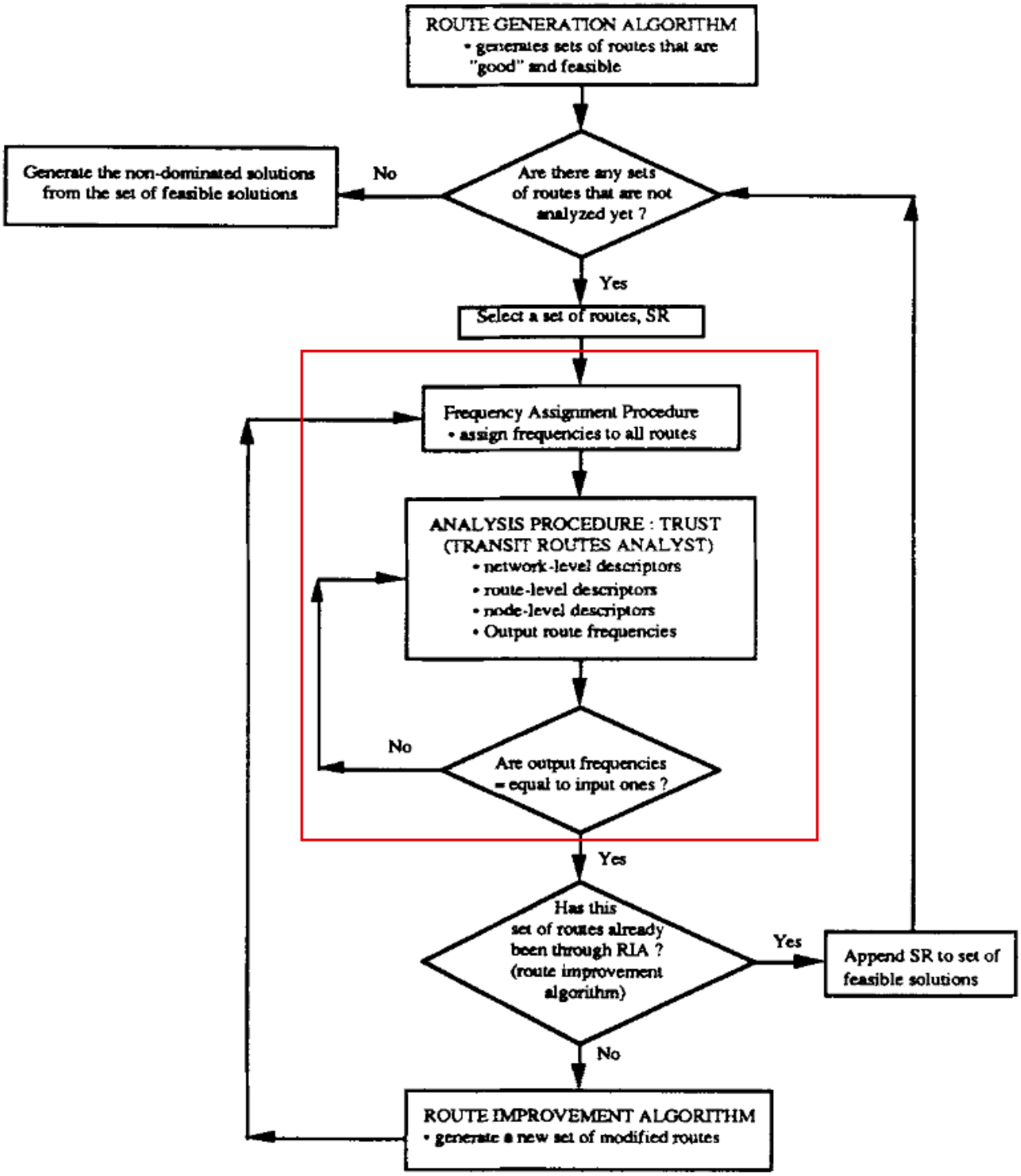

Figure 4.6: Baaj and Mahmassani's solution approach (extracted from [43])

Table 4.3: Mandl - Parameter configuration

	Max. fleet size	**Frequencies**
OPT1	105	$\{0.1, 0.3, 0.4, 0.6, 0.8, 1.0, 1.15\}$
OPT2	110	”
OPT3	120	”
OPT4	140	”
OPT5	160	”
OPT6	200	”
OPT7	105	$\{0.1, 0.2, 0.3, 0.6, 0.8, 1.15, 1.20\}$
OPT8	110	”
OPT9	130	”
OPT10	140	”
OPT11	160	”
OPT12	200	”

the calculated flows in an straightforward manner, so transfer measures are not included in our results.

In Table 4.4, we show the solutions obtained by Baaj and Mahmassani (BM1, BM2, and BM3), those reported by Mandl (Mandl1 and Mandl2), and those obtained by running our proposed formulation (denoted as OPT) with different sets of parameters, as shown in Table 4.3. The parameters differ in their fleet sizes and the set of frequencies available. They were chosen to achieve a reasonably good diversity of solutions, with different tradeoffs regarding the operator's objective (minimizing the total fleet size) and the users' objective (minimizing their total travel times). Still, the two sets of frequencies are not that different. This is explained by the nature of the solutions reported in Baaj and Mahmassani, and Mandl, which are concentrated in a rather specific region of the Pareto set due to the way they assign frequencies to the different lines. Thus, to compare the various solutions, we constrained the frequency set to the given range, which is quite large, to the point that some of the frequencies involved are too high to be practicable in real urban contexts.

Table 4.4 shows the solutions obtained after executing 12 runs of formulation (3.48 - 3.75). We show the total travel time (in-vehicle travel time plus waiting time), in-vehicle travel time, waiting time, fleet size, percentage of demand satisfied by direct travels (“Direct” column) and by one transfer (“1-T” column), the average mean utilization across all lines (in percentages, the “U” column), and finally the execution time that each run took. The formulation consisted of 174.672 variables (84.200 of them binary) and 261.373 constraints. All runs except for OPT11 and OPT12 were done using CPLEX 12 in a Core i7 computer of 3.4 GHz with 16 GB of RAM. Both OPT11 and OPT12 reached the memory limit, and the executions were aborted, so we retried them on a Core i9 computer with 64 GB of RAM.

The mean utilization U is defined for each line $l \in L$ as

$$U = \frac{\phi}{\sum_{f \in \Theta} y_{lf} \theta_f \omega} \tag{4.1}$$

Table 4.4: Mandl - Result comparision

	Total	**In-vehicle**	**Waiting**	**Buses**	**Direct**	**1-T**	**U(%)**	**t**
BM1	3150	2801	349	89.3	78.61	21.39	-	2.65 s
BM2	3250	2818	432	76.9	79.96	20.04	-	3.08 s
BM3	3468	3006	462	82.2	80.99	19.01	-	2.13 s
Mandl1	3768	3515	253	116	68.85	31.15	-	-
Mandl2	3258	2956	302	99.3	69.94	29.93	-	-
OPT1	3608	3041	567	104.1	-	-	65.2	1h 19m
OPT2	3541	3041	500	108.1	-	-	51.1	42m
OPT3	3461	3041	419	118.7	-	-	45	28m
OPT4	3390	3041	349	139.9	-	-	36.5	1h 17m
OPT5	3352	3041	310	159.9	-	-	30.2	1h 25m
OPT6	3325	3041	283	188.6	-	-	26.6	1h 18m
OPT7	-	-	-	-	-	-	-	4h 20m
OPT8	-	-	-	-	-	-	-	15h 30m
OPT9	3430	3041	389	128.7	-	-	40.7	1h 40m
OPT10	3397	3041	356	136.3	-	-	35.8	1h 35m
OPT11	3348	3041	307	158.8	-	-	37.7	1h 59m
OPT12	3313	3041	272	196.8	-	-	25.4	1h 15m

where

$$\phi = \frac{\sum_{k \in K} \sum_{a \in A} c_a v_{ak}}{\sum_{a \in A} c_a} \quad \forall\ a \in l \tag{4.2}$$

It is interesting to note that, in the optimal solutions, the in-vehicle travel time remains the same across runs. We identify many reasons for this. First, since our formulation considers a fixed set of lines, as opposed to the algorithms by Baaj and Mahmassani and Mandl, which attempt to tackle the more general Transit Network Design Problem, it is expected that the optimal solutions found would have fewer passenger flow variations as the itineraries for every line remain fixed, causing the demand to be captive of some selected street travel paths. Still, even when the itineraries are fixed, the passenger flow could undergo a redistribution as the various frequencies are assigned to different lines. This would represent demand that prefers to board a different line than initially due to, for example, lower waiting times. This does happen in the solutions encountered by our formulation, but very seldom. To understand the causes for this, we note that there is very little overlap between the lines reported by Mandl in terms of the streets they travel through, which means that most of the demand is captive of just a single line. Even when overlap exists, it only happens through very short paths using the same street segments as the alternative line, which would then yield the exact same travel times even though a different line is serving the demand. Furthermore, the lines that drive most of the demand will usually hold the faster frequencies in order to minimize the travel times, subject to the usual constraints such as the fleet size. Indeed, we note a monotonic increase across the solutions in the frequencies assigned to each line every time we allow more buses to operate. This results in the same set of lines getting the most rapid frequencies as the

fleet size constraint is relaxed. Moreover, a significant part of the demand is concentrated on very few nodes, which increases the pressure of the lines that serve that portion of the street network to get the highest frequencies. Finally, the fact that some of the frequencies are much higher than others makes the frequency share rule less relevant when it comes to distributing the flows, something that also speaks about the artificial nature of the case study under discussion, as mentioned in Section 4.2

The solver could not find a solution for some instances (OPT7 and OPT8). In those cases, the run was stopped after the specified amount of time passed, with the solver informing that no integer feasible solution was found.

As discussed in Subsection 4.2.1, several differences exist among the published algorithms and the formulation proposed in this work. The most important aspects that diverge are the different line itineraries and assignment sub-models. This explains the discrepancies, particularly in the in-vehicle travel times, waiting times, and computed fleet sizes. Nevertheless, the measures are comparable in magnitude for every reported metric.

4.2.3 Analysis of the solutions

In this section, we study the model's behavior when applied to Mandl's instance. We single out some of the OPT solutions presenting interesting properties (i.e., those where the capacity constraint is more active) to analyze their nature in detail with the help of various reported metrics and a visual inspection of the underlying transport network.

From Figure 4.7, we can observe that a wide range of trade-off levels is achievable by varying the fleet size restriction (parameter B). The monotonic tendency between the total travel time and the fleet size is apparent in the case of the optimal solutions since only the frequencies assigned to the lines can change while the line itineraries remain fixed, and since alternative lines for the same demand pairs usually overlap through the same street paths. In the case of the solutions obtained by other means, the monotonic trend prevails but is not guaranteed since the line itineraries are also subject to change. In general, this trade-off depends strongly on the number of lines considered and on their frequencies: solutions reporting lower total travel times (and thus deemed attractive to the users, i.e. solutions OPT5, OPT12) usually require more lines and higher frequencies, while solutions with a low cost for the operators (resulting in longer total travel times, i.e. OPT1, OPT9) typically present lower frequencies. It can also be noted that solutions that favor keeping operator costs low (by setting low frequencies) incur larger mean utilization rates of the lines (more crowded buses on average), which is to be expected.

In general, the mean utilization of buses is lower than 100% for all encountered solutions; higher values correspond to solutions requiring smaller fleet sizes. Notice that due to the bus capacity constraint, the mean utilization can never be above 100%.

In Table 4.5, detailed information is shown for two selected solutions that differ in the domain of frequencies considered and the maximum number of buses allowed to operate. The solutions were chosen because they incur the lowest costs from the operators' point of view and the parameters considered. In this case, we report the maximum and minimum waiting times individual users experience (Δt_{max} and Δt_{min}, respectively) and other line-specific measures such as the line capacity (defined in 3.25), the critical flow, the line utilization (U, defined in Equation 4.1), and the line headway ($1/f$, the inverse of the line

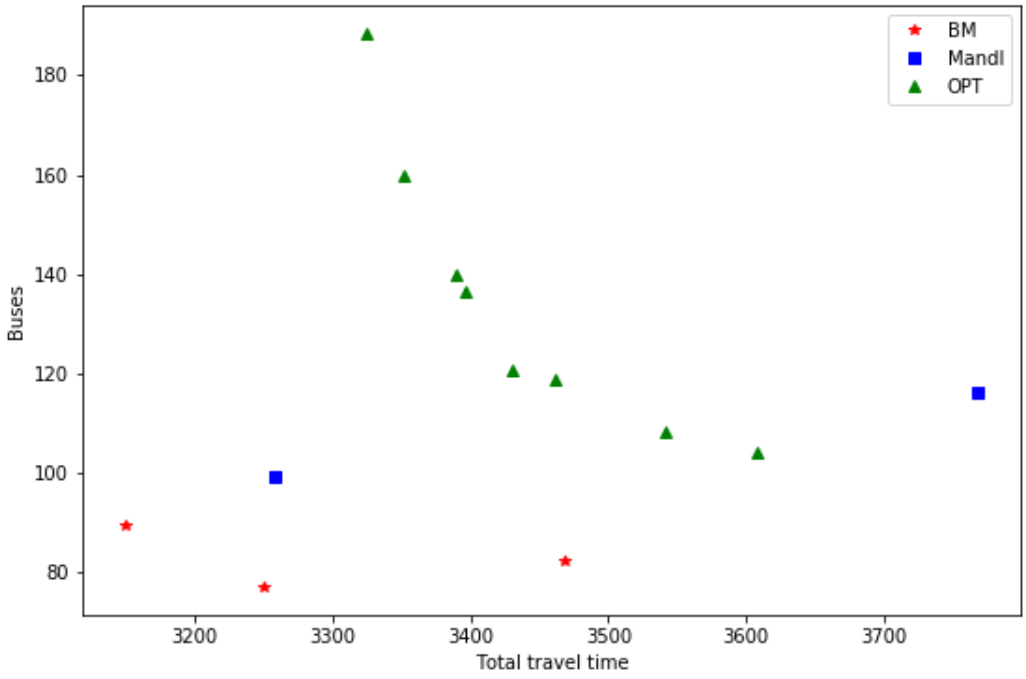

Figure 4.7: Mandl results

frequency). The critical flow of a given line l is defined as ϕ_l^*:

$$\phi_l^* = \max_{a \in A_l^T} \sum_{k \in K} v_{ak} \tag{4.3}$$

Since w_{nk} includes not only the waiting times but also a factor representing the actual flow traveling through it, Δt is calculated as follows:

$$\Delta t = \frac{w_{nk}}{\sum\limits_{a \in A_n^+} v_{ak}} \quad \forall n \in N^P, k \in K \tag{4.4}$$

It is interesting to note that some of the lines exhibit buses very close to the congestion point, in the sense of the passenger load with respect to the capacity of the buses. This is true for all lines in OPT1, and for line 1 in OPT9. In the first example, even a slight decrease in the frequencies for some of the lines will most likely invalidate the solutions due to the bus capacity constraint, while in the second case, if there are alternative trajectories for the demand pairs covered by line 1, then the demand will spread among the lines covering those alternative paths (probably at the expense of higher total travel times). Despite that, the line utilization remained well below 100%, suggesting that the congestion effects are local to very specific parts of the underlying network. In that case, an alternative set of line trajectories to the ones considered could alleviate the issue. Indeed, that seems

Table 4.5: Mandl - Additional measures

	Δt_{max}	Δt_{min}	**Line**	**Capacity**	**Critical flow**	**U(%)**	$1/f$
OPT1	10	0.54	1	57.5	56.9	57	0.87
			2	20	18.8	66	2.5
			3	15	13	59	3.3
			4	5	4.6	79	10
OPT9	3.33	0.43	1	57.5	56.9	57	0.87
			2	30	18.8	44	1.7
			3	30	13.9	29	1.7
			4	15	5.6	34	3.3

to be the case, and can probably explain the improved solutions found by methods that seek to optimize line trajectories in addition to frequencies.

Moreover, we can observe that since the capacity constraint is close to its limit for all the lines in OPT1, this solution probably belongs to one of the extreme points in the optimal Pareto front, leaning towards the region with lower costs for the operators (for a Pareto front calculated with the given domain of frequencies and a variable fleet size). Indeed, experimenting with lower fleet sizes proved this assumption, and no solution was found below $B = 105$. We note here that this is one of the benefits of using exact models: they can prove (guarantee) the nonexistence of feasible solutions, whereas, in heuristics and metaheuristics approaches, it is generally not possible to rule out their existence just because the resolution method was not able to find one.

We also note that even with very high frequencies (i.e., notice the headway for line 1), the critical flow is very close to turning the solution into an infeasible one. The utilization of very high frequencies seems to be the case for the solutions reported by the other two algorithms.

In Figure 4.8, we can observe the arcs where the critical flows occur for each of the lines in OPT1 (dashed lines). Due to the set of lines considered, demand originating at the upper side of the graph that has as a destination one of the nodes on the lower side has very few travel alternatives (and vice-versa). Indeed, for most demand pairs that fall into this category, some combination using lines 1 and 4 is required, thus making node 10 a very affluent one. It would be reasonable to think that with the addition of new lines following paths connecting each region through node 12, the congestion at that point in the network might be avoided. After some tests, this proved to be a false assumption. As mentioned before, OPT1 lies in one of the extreme points of the Pareto Set, corresponding to the lowest cost for the operators. No other feasible solutions were found for $B < 105$: either the fleet constraint was invalidated, or part of the demand was unmet. The addition of any new line to the model, in combination with the extremely high frequencies needed to compare the different solutions, would result in more buses, thus making the fleet restriction infeasible (OPT1 requires a total fleet of 104.1 buses while the restriction caps that total at 105).

Finally, we note that the maximum waiting time for individual users is much higher than that of those with the lowest value. This is particularly true in OPT1, where the least fortunate users must wait for about 1849% more time than the most privileged ones.

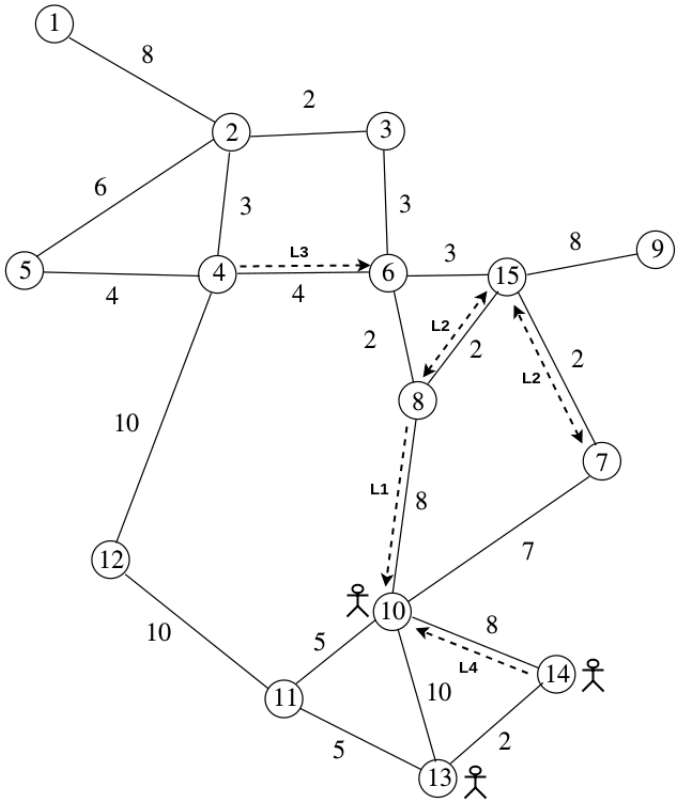

Figure 4.8: Mandl - Critical flows and maximum waiting times

In Figure 4.8, a passenger symbol depicts the nodes where users must wait most. They correspond to users waiting to board buses belonging to line 4, which is to be expected since this is the line that operates with the lowest frequency in the solution. This suggests that it might be worthwhile to consider the addition of a new constraint that can ensure a maximum waiting time for all of the users of the transit system. Even if the overall waiting times are somewhat optimized due to their inclusion in the objective function, this situation could be regarded as an unacceptable level of service. We run more experiments considering a maximum waiting time constraint and analyze its implications for the model in the following section.

4.3 Adding the maximum waiting time constraint

Formulation (3.48 - 3.75) is going to favor solutions that incur overall short waiting times but can still impose arbitrary long waiting times for a non controlled percentage of the users. Even when the system achieves a good overall performance from the users' viewpoint, it will not be acceptable for the users of the affected OD pairs. To address this problem, we can add a maximum waiting time constraint. This constraint belongs to the upper-level problem (planner scope) and involves variables of the lower level. It can be expressed as:

$$\sum_{a \in A_n^+} v_{ak} > w_{nk}/\epsilon \quad \forall n \in N^P, k \in K \tag{4.5}$$

where ϵ is the maximum allowable waiting time for any user of the system and w_{nk} is the waiting time multiplied by the amount of demand k at node n, first defined in Equation 3.5. When we divide w_{nk} by the sum of the flows of the boarding arcs leaving node n, we get the total waiting time at the node for the given demand, which is the value we desire to constrain. We also note that parameter ϵ has a very direct interpretation, allowing planners to apply specific politics with regard to the level of quality of the offered services.

Since Equation 4.5 pertains to the upper level problem, it can be added directly to formulation (3.48 - 3.75) as an additional constraint without affecting the Karush Kuhn Tucker reformulation.

In order to address the wide disparity in waiting times shown in Table 4.5 for OPT1, we reran the formulation with the new waiting time constraint and $\epsilon = 10$, which is the maximum waiting time originally reported by the solution. This yields no feasible solution, and it is not hard to understand why. As mentioned earlier, OPT1 lies in one of the extreme points of the Pareto set, corresponding to the operators' lowest cost. The amount of buses required to operate the optimal solution is already very close to the maximum number of buses allowed (104.1 and 105, respectively). If we were to lessen the waiting times for a part of the demand, a real-world scenario would require either to:

1. Assign a higher frequency to the lines serving the affected demand.
2. Change the itineraries of the lines in such a way that it alleviates the waiting times for affected demand.
3. Operate buses with larger capacities.

Option (1) means, in the context of this instance of the problem, adding more buses to the solution, but as explained, the bus fleet constraint is already active in OPT1, to the extent that even adding just one more bus would violate the total fleet size constraint. Option (2) would require us to solve a different problem, in which not only the frequencies but also the line itineraries were susceptible to change, that is, to tackle the General Transit Network Design Problem, whereas option (3) would entail a relaxation of the bus capacity constraint, thus forcing us to consider a different (and a relaxed) instance of the problem.

In conclusion, the interplay between the total fleet size, bus capacities, and the waiting time is such that it's very hard to constrain the maximum waiting times of an already optimal solution. Since the objective of the formulation is to minimize the total travel times, this usually results in activating the total fleet size restriction, as more buses operating over the network typically mean lower waiting times and more capacity overall. When this restriction is close to its upper bound, there's little room for improvement.

Since finding a solution with lesser maximum waiting times that respects the original fleet size constraint was not possible, an alternative is to find the next best solution respecting both. In doing so, in a manner reminiscent of what was done in Subsection 4.1.2, we set up to work with an alternative formulation, where the objective function is to

Table 4.6: Mandl - Result with maximum waiting time constraint

	Total	**In-vehicle**	**Waiting**	**Buses**	Δt_{max}	Δt_{min}	**U(%)**	**t**
OPT1	3608	3041	567	104.1	10	0.54	65.2	1h 19m
OPT_{wt}	3541	3041	500	108.1	3.33	0.54	51.1	13h 31m

minimize the total fleet size, subject to a maximum total travel time and waiting times. By demoting the total travel time to a constraint in formulation (3.48 - 3.75), we can set its upper bound as the optimal value encountered by OPT1. Thus, we seek the minimum number of buses that can operate a solution where the level of service offered to the users is at least as good as in OPT1 (in terms of the total travel times) while providing lesser values of waiting times for the less lucky users. In terms of formulation (3.48 - 3.75), this requires us to turn the original objective function (3.48) into a constraint, using (3.49) as the new objective function while dropping parameter B, and adding the new maximum waiting time constraint (4.5).

In Table 4.6, we show the results after running the proposed alternative formulation with the same set of parameters as for OPT1, where OPT_{wt} represents the new solution found. We show the total travel time, the in-vehicle travel time, the waiting time, the number of buses required by the solution, the maximum and minimum waiting times individual users experience (Δt_{max} and Δt_{min}, respectively), the average mean utilization across all lines (U), and the execution time that each run took (t). We ran the experiment using a Core i9 computer with 64 GB of RAM.

We notice that the new constraint was effective, achieving an improvement of 66.7% in the maximum waiting times at the expense of increasing the fleet by 3.8% (four more buses). Overall waiting times were improved by 12%, while the mean line utilization dropped by 22%.

In Figure 4.9, we can observe the arcs where the critical flow occurs for each of the lines in OPT_{wt} (dashed segments) as well as the zones where users must wait the most (depicted by a passenger symbol).

When compared to OPT1, depicted in Figure 4.8, the critical flows remain for the most part the same, except for the critical flow corresponding to line 1, which reverses its direction, and the critical flow corresponding to line 4, which now goes in both directions. We also note that the original maximum waiting times continue to occur at the same bus stops (nodes 10, 13, and 14), but are now even in magnitude with the waiting times occurring elsewhere for different lines, since the actual time spent waiting has been considerably reduced. In a way, this more uniform distribution of the waiting times throughout the transport network and the different lines can be interpreted as a more "fair" distribution that does not overly penalize a specific subset of the demand at the expense of the rest.

In Table 4.7, we display some additional measures per line. Despite OPT_{wt} incurring a more acceptable maximum waiting time than OPT1, we note that the solutions don't differ as widely in other respects: all the metrics for the first three lines are identical, as they continue to operate using the same set of frequencies. The main change in terms of the lines occurs with line 4, which in OPT_{wt} uses a higher frequency necessary to diminish the waiting time of the more affected users. We showed in Figure 4.8 that the maximum waiting times are local to specific nodes in the network, and the low critical flow and mean

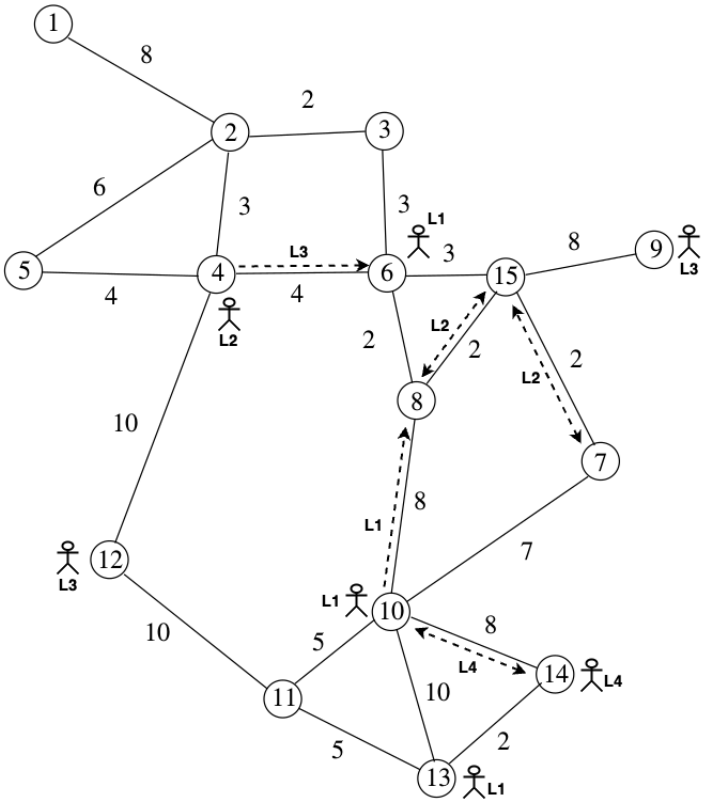

Figure 4.9: Mandl - Critical flows and maximum waiting times when using a maximum waiting time constraint

utilization for line 4 in the new solution confirm this: the frequency required to alleviate the high waiting times at specific stops in the network turns to be wasteful for the rest of the itinerary of the line, which now operates with a much higher capacity than needed for most of its journey. This kind of analysis could form the basis to further explore and fine-tune the set of feasible frequencies available for the instance under consideration. It can also be interpreted as an argument in favor of using continuous variables for the frequencies. The fact that the critical flow for line 4 is close to the one in OPT1 suggests that a frequency higher than the original but lower than the final selected would constitute a better option in terms of line utilization. Alternatively, when there is such disparity in the utilization of the different travel arcs, one approach used in practice is to follow a short-turn line strategy [21], where lines that operate over a very localized part of the network with high frequencies are introduced.

Finally, we note that OPT_{wt} is identical to OPT2, one of the solutions found by using formulation (3.48 - 3.75) without any constraint on waiting times and with a total fleet size constraint of five more buses with respect to instance OPT1. However, in terms of computational performance, while OPT2 took just 42 minutes to execute, OPT_{wt} required 13 hours and 31 minutes. The increase in the execution time can be explained by the

Table 4.7: Mandl - Additional measures with the maximum waiting time constraint

	Δt_{max}	Δt_{min}	**Line**	**Capacity**	**Critical flow**	**U(%)**	$1/f$
OPT1	10	0.54	1	57.5	56.9	57	0.87
			2	20	18.8	66	2.5
			3	15	13	59	3.3
			4	5	4.6	79	10
OPT_{wt}	3.33	0.54	1	57.5	56.9	57	0.87
			2	20	18.8	66	2.5
			3	15	13	59	3.3
			4	15	5.6	34	3.3

addition of the new waiting time constraint, but it can also be due to treating the total travel time as a constraint, or by the promotion of the total fleet size as the objective to be optimized. In any case, this suggests that increasing the total fleet size in small amounts in formulation (3.48 - 3.75) might achieve an equally good tradeoff with regards to the waiting times as using an explicit maximum waiting time constraint.

4.3.1 Main findings

In this section, we proved the efficacy of using the maximum waiting time restriction to achieve a more uniform and fair situation for users of a transportation system. In particular, we were able to improve the maximum waiting times by a considerable percentage while only adding four more buses to the final solution. In doing so, we identified several challenges:

- Adding constraint (4.5) directly to formulation (3.48 - 3.75) results in no feasible solution. In order to leverage the waiting time constraint, we had to turn the total fleet size into the objective to be optimized. The interplay between fleet size, bus capacities, and the minimization of total travel times leaves little room for a maximum waiting time constraint, especially when only the frequencies, not the line itineraries, are subject to change. This could be interpreted as an argument for optimizing both line itineraries and frequencies in a single formulation instead of in a two-step process. Still, operational and planning concerns might prevent the alteration of line itineraries on a more tactical basis, as is often the case for the frequencies.

- Using a model that explicitly avoids congestion results in a rigid setting of frequencies due to the strict nature of the capacity constraint. As a more flexible approach, assignment models that explicitly consider congestion effects, i.e., by considering the influence of the travel flows over the waiting times, might provide a helpful relaxation, giving room for more feasible solutions at the expense of congestion in the network. To be able to analyze a solution, even if it presents prohibitive levels of congestion, is better than having no solution at all, which would leave planners unable to identify which sectors of a transportation network are problematic and could benefit from different measures, such as changes in terms of infrastructure, line itineraries, or the frequency of the lines. Still, we notice that an explicit modeling

of congestion would not necessarily result in finding more feasible solutions when adding constraints such as (4.5); the interplay between the fleet size, bus capacities, and total travel times does not fundamentally change when trying to alleviate high waiting times.

- Near-congestion events often occur in very localized points of the network. Increasing the frequencies of the lines that serve those troubled segments might result in buses with low mean utilization, as previously shown. This suggests that alternative approaches to alleviate congestion might be worth considering, such as models treating frequencies as continuous instead of discrete variables, or the possibility of operating temporary lines over specific saturated zones of the street network.

4.4 Rivera test case

To explore up to which size, in terms of the underlying graph, instances are solvable to optimality, a real test case is considered. It corresponds to the city of Rivera, located in the north of Uruguay and close to the border with Brazil. The case construction reflects the real aspects of an actual city's infrastructure and demand requirements with fidelity. Moreover, the size of Rivera, in terms of the underlying graph, is much larger than Mandl's. Rivera, a city with a population of approximately 65000 inhabitants, has an infrastructure graph of 84 nodes and 143 arcs and a demand matrix consisting of 378 non-null entries. There are several solutions published in the literature obtained by approximate methods, such as metaheuristics [2, 4, 56], and one work where an exact model for frequency optimization is used, but no solution could be obtained [55].

4.4.1 Experiments

Considering the existing publications and the characteristics of the actual solution implemented in Rivera, we establish $\omega = 42$ and $F = (1/60, 1/40, 1/30, 1/20)$ as the set of feasible frequencies. The solution implemented in Rivera consists of 13 bus lines. In terms of our formulation, this translated to 4.576.920 variables (2.198.544 binary) and 6.487.269 constraints. All the experiments in this section were run using a Core i9 computer with 64 GB of RAM. Unfortunately, the CPLEX solver ran out of memory after a couple of iterations, and no solution was found.

Taking into account the execution times required for running the Mandl test case, it is not surprising that a significant increase in the size of the graph poses problems to the solver. Some of the possible causes are:

- As mentioned in [55], the discretization of the domain of frequencies enlarges the set of boarding arcs (and thus the corresponding number of flow variables v) to a factor equal to the size of the set Θ of possible frequencies. It also enlarges the variable space due to the addition of the new binary variables y, which turns the formulation into a mixed integer one.

- Some limited experiments suggest that a large portion of the search tree has to be enumerated when substituting the lower level problem by means of the Karush Kuhn

Tucker conditions in combination with the linearization approach using the big-M method [7].

Chapter 5

Conclusions and further research

We have studied several models in the literature on frequency optimization in public transportation systems, with a particular focus on those that incorporate the characteristics deemed more relevant in the context of this work. The main contributions of this thesis can be succinctly stated as:

- A mathematical programming formulation that explicitly models user behavior (by means of an assignment sub-model) while including bus capacities. This allows us to improve the realism of existing models by considering passenger behavior, bus capacities, and user waiting time at bus stops, which is, under this work's hypothesis, the most relevant part of the total travel time.
- A thorough study of the bilevel nature of frequency optimization in public transportation systems and how the inclusion of different aspects of the problem affects this nature.
- A thorough study of the proposed model behavior, allowing us to identify the advantages and limitations of using a strict capacity approach.

The contributions related to the previous items are supported by the current state of the art in frequency optimization. To the best of our knowledge, the simultaneous inclusion of all of the aforementioned aspects in a single mathematical programming formulation has not been studied. Other elements supporting the merits of the proposed formulation include the broad set of experiments executed, the numerical results obtained using a case study taken from existing publications, and the comparison of the results with results coming from reference studies.

The detailed inspection of several measures of the optimal solutions obtained through different experiments, and the visual assessment of the effects of congestion when contemplating bus capacities, contribute to raising awareness of the importance of bilevel approaches when attempting to solve the frequency optimization problem.

5.1 Bilevel nature

Models that are based on the assumption that planners can assign passengers to lines are common in the literature [3, 13, 65]. Those that consider an explicit assignment sub-model

to reflect user behavior usually do not consider bus capacities [41, 55], nor the waiting time of the users [40]. Since the level of service perceived by the users is one of the most important factors in measuring the performance of a public transportation system, the previous characteristics are of great importance in achieving models with realistic assumptions that can be applied to real scenarios.

The bilevel nature of the problem was explored by applying variants of the formulation to the same instances. The results suggest that a genuine bilevel approach should be considered whenever bus capacities are contemplated, since uncapacitated models can produce solutions that are not appropriate in contexts where the transit system is operating over its capacity. Moreover, by using a test case corresponding to an actual city, it was possible to explore some underlying issues that arise whenever bus capacities and different fleet sizes are considered. By studying measures such as critical flows, line capacities, and maximum waiting times, with the aid of visual inspection, problematic sectors of the underlying line network can be quickly identified. This permits a more thorough discussion of the issues that could arise in real-life contexts and helps devise alternative solutions.

Extensions and alternatives to the proposed formulation were explored to consider different scenarios. An alternative formulation was proposed, capable of determining the minimum fleet size required to operate in order to meet a certain level of service from the users' point of view. By adding new constraints such as maximum waiting times at the bus stops, the flexibility of the proposed model and their impact on the level of service perceived by the users were also analyzed.

5.2 Mathematical formulation

We proposed a new bilevel formulation based on the model presented in [55]. Our model considers individual passenger route choice using an assignment model [72] while also including the waiting time of the users and bus capacities when measuring the system's performance. This makes the model more adequate for being applied in real-case scenarios. We derived a mixed integer linear programming (MILP) formulation, equivalent to the bilevel one, and susceptible of being solved by common solvers using standard MILP techniques.

To study the scalability of the method, we applied the formulation to instances of small to medium sizes. This enabled a detailed comparison with results published in the literature. Several measures extracted from optimal solutions were presented, which can help decision-makers characterize their quality with respect to the interests of both users and operators.

Finally, we have also tested extensions of the model by including and excluding certain constraints in order to study their impact on the mathematical structure of the problem.

5.3 Experiments and application to real cases

This section presents the main conclusions and challenges when performing the computational tests using the proposed mathematical formulation. We also conclude on the feasibility of applying the methodology and the formulation to cases corresponding to actual cities.

5.3.1 Experiments

The main challenge when evaluating the formulation is the lack of comparable results in the literature. In order to compare in a meaningful way two different approaches to frequency optimization, careful consideration must be given to the several aspects outlined in Subsection 4.2.1. In particular, many variants of published models exist, but even when the same sets of variables and constraints are considered, differences usually arise in the assignment sub-models representing the behavior of the users. Another difficulty is the lack of standard test cases and available data in general.

To partially overcome these issues, we used a test case proposed by Mandl, where the data is available, and results have been published in several works [43, 53]. No information was provided regarding the construction of the use case, so it is not possible to map the graph data with the reality of an actual network infrastructure.

To explore the scalability of the proposed methodology, we used a case study corresponding to the actual city of Rivera [56]. The construction process is well documented in this case, and significant effort was devoted to determining a demand matrix that closely resembled the demand requirements of users of the transport services at the moment of writing. From a size perspective, the case study of Rivera is significantly bigger than that of Mandl, something desirable since the main objective was to determine up to which size instances are solvable by optimality using the proposed formulation. Unfortunately, the solver had difficulties when trying to solve the instance, which suggests that cases of comparable size are still out of reach of exact techniques.

The proposed formulation represented the behavior of users with the optimal strategies assignment model [72]. Assignment models are a fundamental part of any model or algorithm attempting to solve the frequency optimization problem; the choice of the assignment model significantly impacts the realism, validity, and interpretation of the solutions. Since this component is a complex model on its own, it also greatly impacts the performance of solving algorithms. Another assignment model mentioned in the context of this work is that of Baaj and Mahmassani [43], which can be regarded as similar to optimal strategies, although it does not consist of an explicit mathematical formulation, something that prevents its use as a sub-model embedded in explicit formulations.

5.3.2 Application to real cases

There are several possible applications of the proposed formulation in the design of a public transportation system.

A given frequency setting is needed in the context of strategic planning when an initial set of lines and their itineraries must be defined. In that context, frequencies are required to evaluate the quality of the lines both from the users' point of view (since frequencies affect the waiting times) as well as from the operators' point of view (since frequencies strongly determine the total number of buses that they need to provide). Moreover, frequency setting usually occurs during tactical planning, in response to updated demand requirements, among other dynamic factors.

Adjusting the frequencies is also a straightforward approach to achieving different trade-offs between the operators and user objectives.

Most of the published results in the literature proposing heuristic or metaheuristic approaches lack a way of objectively comparing their algorithms' performance. By leveraging the proposed mathematical programming formulation or other exact methods with optimality guarantees, it is possible to better assess the quality of the solutions encountered by said algorithms, for example, by calculating an optimality gap over selected instances.

Finally, it is possible to perform alternative experiments with only slight changes to the formulation. In this thesis, we have used a derived formulation to estimate the fleet size required to ensure a certain level of service from the users' point of view. We have also devised an alternative formulation that includes a maximum waiting time constraint to provide a more consistent and fair level of service to the users when waiting at the different bus stops.

5.4 Further research

We note that all variants of the bilevel model discussed in this work maintain the DCLB structure, enabling the application of exact resolution methods. However, the existing (general purpose) methods for this kind of bilevel problems [7, 9, 31] do not necessarily handle models with many variables and constraints, as is the case for 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 [40]. Metaheuristic techniques may also aid in finding good solutions to solve the transit frequency optimization problem. The *Tabu Search* [39] based metaheuristic presented in [55] 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 explicitly deal with bilevel problems. A good survey can be found in [73].

Regarding capacitated models, a formal criterion for switching between uncongested and congested frequency optimization models is desirable to establish. The capacity of the buses is modeled by either assuming that the planner ensures sufficient capacity on the lines that the users want to use or by the explicit treatment of a congested system, usually using an equilibrium assignment sub-model. Several practical aspects of reality determine whether it is feasible to operate a not congested system. Still, to the best of our knowledge, there are no published works exploring and comparing these different approaches over the same instances.

To improve the model's realism further, new constraints could be considered. Some constraints, such as imposing street capacities, can be easily added to the proposed formulation without significant changes.

In the current formulation, all demand pairs must be covered by one or some combination of lines, a situation imposed by the flow conservation constraint. In some scenarios, particularly in big cities, this might be deemed unrealistic, and thus demand covering constraints could be helpful. Demand covering constraints consider the case where only a certain percentage of the demand needs to be satisfied. They can also be extended to consider upper and lower limits on the total amount of demand covered directly or by one or more transfers. The model proposed in this thesis supports transfers, but does not penalize them explicitly, since doing so would increase the complexity of the current formulation. Still, considering some form of penalty could be worthwhile in contexts where

transfers are viewed in such a negative light as to significantly influence the perception of the level of service that the users experience, to the point of influencing user behavior and, thus, the underlying assignment model.

Certain assumptions mentioned in Section 1.1 limited the scope of the proposed models. Including aspects such as the influence of fares charged for the usage of the services, the availability of advanced traveler information systems, the consideration of transfers, or the disregard of waiting times in high-frequency contexts would significantly affect the structure of the proposed formulation, especially with respect to the assignment sub-model, since those aspects influence the behavior of the users of the system.

Finally, we note that finding feasible solutions was hard in some instances. A similar conclusion was reached in [19], where the author states that a bit of congestion is not unreasonable. This motivates her to devise an alternative formulation, where the fleet size constraint is instead treated as a penalized factor of the objective function. The formulation proposed here could be extended similarly with little effort.

Appendix A

TRUST assignment model

The TRUST assignment model used by Baaj and Mahmassani in [43] (and described in detail in [42]) has similar characteristics to optimal strategies [72], which is used in the context of this work and further explained in Section 3.3). When compared, the salient characteristics of TRUST are:

- Transfers are considered in the behavior of the users when selecting lines.
- Congestion effects are not considered.
- A frequency share rule is used when multiple lines are capable of serving the demand.
- Total travel time minimization is performed in a heuristic way instead of using exact techniques with optimality guarantees.

While transfers are allowed in optimal strategies, they are implicitly handled, so it is not possible to quantify them within the model. On the other hand, TRUST utilizes the number of transfers a user must take as the primary criterion when choosing among different lines. If the minimum number of transfers for a given OD pair can be achieved by more than one strategy, the strategy that minimizes the total travel time is selected, as in the optimal strategies case.

Congestion is not modeled in TRUST. Optimal strategies does not consider congestion either: the cost of the street edges is thus fixed, and solutions that lack enough bus capacity are discarded. This means that operators need to guarantee sufficient capacity when, for example, assigning frequencies.

We also note that the frequency share rule used by TRUST is equal to the one depicted in Equation 3.4, while the waiting time expression is similar to the one depicted in Equation 3.1 with $\beta = 1/2$.

Appendix B

Mandl test case clarification

While validating the proposed mathematical formulation using the test case of Mandl, we noticed some inconsistencies in the way the results were reported in [43]. In particular, the original article states that the origin-destination matrix is expressed in travel trips per day, but the reported results are then presented and utilized as if they were travel trips per hour. While, from a performance point of view, this does not prevent or invalidate the comparison of results, it does affect the interpretation of the results from the perspective of a real public transport system. Moreover, the demand seems to be treated on a per-minute basis when calculating capacities and user flows but on a per-hour basis when computing the different components of the total travel time.

In order to present consistent results, in the current work we divide the travel times (including waiting times), as published in the original article, by 60. In this manner, the different metrics of each solution are reported using minutes as the unit.

Bibliography

[1] OR library. `http://people.brunel.ac.uk/~mastjjb/jeb/info.html`.

[2] Andrés Américo, Fernando Martínez, Antonio Mauttone, and María E. Urquhart. Multi-objective evolutionary algorithm for the transit network design problem. In *ICORD VI International Conference on Operational Research for Development*, Fortaleza, Brazil, 2007.

[3] Anita Schöbel and Susanne Scholl. Line planning with minimal transfers. In Proceedings of the fifth Dagstuhl seminar workshop on algorithmic methods and models for optimization of railways. *Transportation Science*, 06901, 2006.

[4] Agustín Arizti. *Algoritmos eficientes para el problema del diseño óptimo de redes de transporte público.* Bachelor's thesis, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, 2014.

[5] José M. Arroyo. Bilevel programming applied to power system vulnerability analysis under multiple contingencies. *IET Generation, Transmission and Distribution*, 4(2):178–190, 2010.

[6] Jonathan Bard and James T. Moore. A branch and bound algorithm for the bilevel programming problem. *Siam Journal on Scientific and Statistical Computing*, 11, 1990.

[7] Jonathan F. Bard. *Practical bilevel optimization.* Kluwer, 1998.

[8] Jonathan F. Bard and James E. Falk. An explicit solution to the multi-level programming problem. *Computers and Operations Research*, 9(1):77–100, 1982.

[9] Luis Baringo and Antonio J. Conejo. Transmission and Wind Power Investment. *IEEE Transactions on Power Systems*, 27(2):885–893, 2012.

[10] D. Bertsimas and J.N. Tsitsiklis. *Introduction to Linear Optimization.* Athena Scientific, 1997.

[11] Wayne Bialas and Mark Karwan. On two-level optimization. *IEEE Transactions on Automatic Control*, 27:211–214, 1982.

[12] Christian Blum and Andrea Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys (CSUR)*, 35(3):268–308, 2003.

[13] Ralf Borndörfer and Marika Karbstein. A Direct Connection Approach to Integrated Line Planning and Passenger Routing. In Daniel Delling and Leo Liberti, editors, *12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, volume 25 of *OpenAccess Series in Informatics (OASIcs)*, pages 47–57, Dagstuhl, Germany, 2012. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

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

[15] 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: Methodological*, 40(6):437–459, 2006.

[16] Partha Chakroborty. Genetic algorithms for optimal urban transit network design. *Computer-Aided Civil and Infrastructure Engineering*, 18(3):184–200, 2003.

[17] Esteve Codina, Angel Marin, and Francesc Lopez. Modeling and optimization of frequencies on congested bus lines. In *Proceedings of the 14th international conference on Automatic Control, Modelling Simulation, and Proceedings of the 11th international conference on Microelectronics, Nanoelectronics, Optoelectronics*, pages 104–109, 2012.

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

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

[20] 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.

[21] Cristián Cortés, Sergio Jara-Díaz, and Alejandro Tirachini. Integrating short turning and deadheading in the optimization of transit services. *Transportation Research Part A: Policy and Practice*, 45:419–434, 2011.

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

[23] J. de D. Ortúzar and L. Willumnsen. *Modelling Transport.* John Wiley and Sons, 1996.

[24] Stephan Dempe. *Bilevel Programming - A survey, Chapter 1.* Technical Report TU 2003-11. Bergakademie Freiberg, 2003.

[25] Guy Desaulniers and Mark D. Hickman. Chapter 2: Public Transit. In Cynthia Barnhart and Gilbert Laporte, editors, *Transportation*, volume 14 of *Handbooks in Operations Research and Management Science*, pages 69–127. Elsevier, 2007.

[26] Robert B. Dial. Transit pathfinder algorithm. *Highway Research Record*, 205:67–85, 1967.

[27] Matthias Ehrgott. *Multicriteria Optimization.* Springer-Verlag, Berlin, Heidelberg, 2005.

[28] Matthias Ehrgott and Xavier Gandibleux. Approximative solution methods for multiobjective combinatorial optimization. *TOP*, 12(1):1–63, 2004.

[29] Lang Fan and Christine L. Mumford. A metaheuristic approach to the urban transit routing problem. *Journal of Heuristics*, 16(3):353–372, 2010.

[30] Jörg Fliege and Luis N. Vicente. Multicriteria approach to bilevel optimization. *Journal of Optimization Theory and Applications*, 131(2):209–225, 2006.

[31] 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.

[32] Robert G. Jeroslow. The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming*, 32:146–164, 1985.

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

[34] 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.

[35] Michel Gendreau. *Étude approfondie d'un modéle d'équilibre pour l'affectation des passangers dans les réseaux de transport en commun.* Doctorate thesis on informatics, Université de Montréal, 1984.

[36] Michel Gendreau. An introduction to tabu search. In Fred Glover and Gary A. Kochenberger, editors, *Handbook of Metaheuristics*, pages 37–54. Springer US, 2003.

[37] Guido Gentile, Michael Florian, Younes Hamdouch, Oded Cats, and Agostino Nuzzolo. The theory of transit assignment: Basic modelling frameworks. In *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems: COST Action TU1004 (TransITS)*, pages 287–386. Springer International Publishing, Cham, 2016.

[38] 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.

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

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

[41] Junfei Guan, Hai Yang, and Sumedha C. Wirasinghe. Simultaneous optimization of transit line configuration and passenger line assignment. *Transportation Research Part B: Methodological*, 40(10):885–902, 2006.

[42] M Hadi Baaj and Hani Mahmassani. TRUST: A LISP program for the analysis of transit route configurations. *Transportation Research Record Journal of the Transportation Research Board*, 1283:125–135, 1990.

[43] M Hadi Baaj and Hani Mahmassani. An AI-based approach for transit route system planning and design. *Journal of Advanced Transportation*, 25:187–209, 1991.

[44] Anthony F. Han and Nigel H.M. Wilson. The allocation of buses in heavily utilized networks with overlapping routes. *Transportation Research Part B: Methodological*, 16(3):221–232, 1982.

[45] Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for bilevel linear programming. *SIAM Journal on Scientific and Statistical Computing*, 13:1194–1217, 1992.

[46] Deb Kalyanmoy. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[47] Bahar Y. Kara and Vedat Verter. Designing a road network for hazardous materials transportation. *Transportation Science*, 38(2):188–196, 2004.

[48] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.

[49] J. Enrique Fernández L., Joaquín de Cea Ch., and R. Henry Malbran. Demand responsive urban public transport system design: Methodology and application. *Transportation Research Part A: Policy and Practice*, 42(7):951–972, 2008.

[50] Homero Larrain and Juan Carlos Muñoz. Public transit corridor assignment assuming congestion due to passenger boarding and alighting. *Networks and Spatial Economics*, 8(2):241–256, 2008.

[51] François Legillon, Arnaud Liefooghe, and El-Ghazali Talbi. Cobra: A coevolutionary metaheuristic for bi-level optimization. In El-Ghazali Talbi, editor, *Metaheuristics for Bi-level Optimization*, pages 95–114. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[52] 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: Methodological*, 44(10):1186–1201, 2010.

[53] Christoph E. Mandl. Evaluation and optimization of urban public transportation networks. *European Journal of Operational Research*, 5(6):396–404, 1980.

[54] Yannis Marinakis and Magdalene Marinaki. A bilevel particle swarm optimization algorithm for supply chain management problems. *Studies in Computational Intelligence*, 482:69–93, 2013.

[55] 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.

[56] Antonio Mauttone. Optimización de recorridos y frecuencias en sistemas de transporte público urbano colectivo. Master's thesis, PEDECIBA Informática, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, 2005.

[57] Antonio Mauttone. *Models and algorithms for the optimal design of bus routes in public transportation systems.* PhD thesis, PEDECIBA Informática, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, 2011.

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

[59] Luis N. Vicente and Paul H. Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global Optimization*, 5:291–306, 1994.

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

[61] Otto Anker Nielsen. A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological*, 34(5):377–402, 2000.

[62] Yolanda Noriega and Michael Florian. L'optimisation des fréquences d'un réseau de transport en commun moyennement congestionné. *INFOR: Information Systems and Operational Research*, 41:129–153, 05 2003.

[63] Agostino Nuzzolo. Transit path choice and assignment model approaches. In *Advanced Modeling for Transit Operations and Service Planning*, pages 93–124. Emerald Group Publishing Limited, 2002.

[64] Victor Oduguwa and Rajkumar Roy. Bi-level optimisation using genetic algorithm. In *Proceedings 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS 2002)*, pages 322–327. IEEE, 2002.

[65] Ralf Borndörfer, Martin Grötschel, and Marc E. Pfetsch. A column-generation approach to line planning in public transport. *Transportation Science*, 41(1):123–132, 2007.

[66] K. V. K. Rao, S. Muralidhar, and S. L. Dhingra. Public transport routing and scheduling using genetic algorithms. In *8th International Conference on Computer Aided Scheduling of Public Transport*, pages 21–23. Berling, Germany, 2000.

[67] Mauricio G. C. Resende and Celso C. Ribeiro. Greedy randomized adaptive search procedures. In Fred Glover and Gary A. Kochenberger, editors, *Handbook of Metaheuristics*, pages 219–249. Springer US, Boston, MA, 2003.

[68] 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.

[69] Siv Schéele. A supply model for public transit services. *Transportation Research Part B: Methodological*, 14(1):133–146, 1980.

[70] Anita Schöbel. Line planning in public transportation: models and methods. *OR Spectrum*, 34(3):491–510, 2012.

[71] Heinz Spiess. *On optimal route choice strategies in transit networks.* Pub 283, Centre de Recherche sur les Transports, Université de Montréal, 1983.

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

[73] El-Ghazali Talbi. *Metaheuristics for Bi-level Optimization*, volume 482 of *Studies in Computational Intelligence.* Springer, 2013.

[74] L. Vicente, G. Savard, and J. Judice. Discrete linear bilevel programming problem. *Journal of Optimization Theory and Applications*, 89(3):597–614, 1996.

[75] Selim Yilmaz and Sevil Sen. Chapter 2: Metaheuristic approaches for solving multiobjective optimization problems. In Seyedali Mirjalili and Amir H. Gandomi, editors, *Comprehensive Metaheuristics*, pages 21–48. Academic Press, 2023.

[76] Yu, B and Yao, J. Genetic algorithm for bus frequency optimization. *Journal of Transportation Engineering*, 136(6):576–583, 2010.

[77] Fang Zhao and Xiaogang Zeng. Optimization of transit route network, vehicle headways and timetables for large-scale transit networks. *European Journal of Operational Research*, 186(2):841–855, 2008.