Dear author,

Please note that changes made in the online proofing system will be added to the article before publication but are not reflected in this PDF.

We also ask that this file not be used for submitting corrections.
Integrated approach to network design and frequency setting problem in railway rapid transit systems

Francisco López-Ramos¹,⁎, Esteve Codina², Ángel Marín³, Armando Guarnaschelli⁴

¹ Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus, Eugenio Garza Sada 2501, 64849 Monterrey, Mexico
² Universitat Politècnica de Catalunya, FIB, Campus Nord, Building C5, 08034 Barcelona, Spain
³ Universidad Politécnica de Madrid, ETS Aeronáuticos, Plaza Cardenal Cisneros, 3, 28040 Madrid, Spain
⁴ Pontificia Universidad Católica de Valparaíso, ITRA, 2147 Brasil Avenue, 2362804 Valparaíso, Chile

ARTICLE INFO

Article history:
Received 16 December 2015
Revised 6 October 2016
Accepted 7 December 2016
Available online xxx

Keywords:
Railway rapid transit
Network design
Working network
Frequency setting
Express service design
Lexicographic goal programming

ABSTRACT

This work presents an optimization-based approach to simultaneously solve the Network Design and the Frequency Setting phases on the context of railway rapid transit networks. The Network Design phase allows expanding existing networks as well as building new ones from scratch, considering infrastructure costs. In the Frequency Setting phase, local and/or express services are established considering transportation resource capacities and operation costs. Integrated approaches to these phases improve the transit planning process. Nevertheless, this integration is challenging both at modeling and computational effort to obtain solutions. In this work, a Lexicographic Goal Programming problem modeling this integration is introduced, together with a solving strategy. A solution to the problem is obtained by first applying a Corridor Generation Algorithm and then a Line Splitting Algorithm to deal with multiple line construction. Two case studies are used for validation, including the Seville and Santiago de Chile rapid transit networks. Detailed solution reports are shown and discussed. Conclusions and future research directions are given.

© 2016 Published by Elsevier Ltd.

1. Introduction

Population growth in cities has led to traffic congestion. To alleviate this effect, transport agencies have designed Rapid Transit Systems. These systems are continuously revised to account for changes in passenger demand.

The overall transit planning process can be divided into the following phases [12]: Transit Network Design (TND), Transit Network Frequency Setting (TNFS), transit network timetabling, vehicle scheduling, and crew scheduling and rostering. Strategic and tactical decisions are taken in the first two phases. Because of the high costs of construction and exploitation of railway transit networks, it is important to optimize every strategic and tactical decision. The TND phase may build from scratch or expand the infrastructure of a rapid transit network (i.e., stations and stretches), considering budgetary restrictions and coverage demand satisfaction. Having determined the new infrastructure of the network, the TNFS sets the line frequencies and the number of vehicles needed to satisfy the passenger trip requirements at reasonable operative costs and not exceeding the capacities of the planning resources (i.e., the number of passengers per vehicle, the number of available vehicles, the maximum stretch frequency, and among others).

In current practice, TND and TNFS phases are solved separately as the infrastructure of the rapid transit network is considered as a stable component contrary to line frequencies which are treated as a flexible component [3]. However, it is admitted that to know how the infrastructure will be used by passengers, an assignment of passengers to lines is required. This assignment requires in turn the definition of lines and frequencies operating on them. Clearly, not considering lines and frequencies in the TND phase makes unrealistic estimates on passenger volumes at this early stage. Therefore, solving separately TND and TNFS phases is an approach that may lead the system to operate inefficiently because the new infrastructure of the rapid transit network is determined without considering the capacities of the planning resources. In other words, the expected amount of demand to be covered in the TND phase can be overestimated because when setting the line frequencies in the TNFS phase, it might not have enough vehicle capacity to fill that demand or the needed line frequencies exceed the stretch capacities. The need for new approaches integrating TND and TNFS phases is acknowledged in [2], and recently some works have explored this integration and showed promising results [4–6] improving substantially rapid transit systems.

⁎ Corresponding author.
E-mail address: francisco.lopez.rt@itesm.mx (F. López-Ramos).

http://dx.doi.org/10.1016/j.cor.2016.12.006
0305-0548/© 2016 Published by Elsevier Ltd.

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
Nevertheless, integrated models impose two great challenges. On the one hand, these models need to correctly represent the underlying structural properties of the TND and TNFS phases and allow a seamless integration. This means that models need to evaluate every trade-off regarding the decisions associated with each phase, and capture the impact of decisions in the TND phase on decisions in the TNFS phase, and vice-versa. On the other hand, the complexity of these models increases because of the integration. Consequently, customized solution algorithms and strategies are required.

In this work, the integration of TND and TNFS phases is achieved by means of an optimization model formulated as a mixed-integer linear goal-programming problem. Because of the size of real-world rapid transit networks, and the highly constrained set of design and frequency decisions, solving the model directly with an exact Branch & Bound optimizer, such as CPLEX, is impractical. Meta-heuristic approaches such as Genetic Algorithms (GA), Simulated Annealing (SA) and Tabu Search (TS), among others, have been used to provide a set of feasible solutions in reasonable time [5-9]. These methods are enough flexible to be adapted to a wide range of optimization models with different objective functions and constraints. However, this flexibility does not allow taking advantage of the mathematical structure of the model to explore the solution space in a way that complexity is minimized. In [7], the performances of GA, SA, TS, among other meta-heuristics, is examined showing that the solving times and the number of algorithmic steps are still strongly affected by the combinatorial nature of problems solving TND & TNFS phases, especially in the number of new lines to be constructed and the enumeration of all possible feasible line traces (corridors).

To deal with these complexity issues, a 2-stage methodology is proposed. The first stage employs a Corridor Generation Algorithm (CGA), whereas the second one uses a mathheuristic (mathematical programming based heuristic) called the Line Splitting Algorithm (LSA). The CGA determines the pool of line traces that can be assigned to the new lines during the optimization process. The size of this pool is limited by infrastructure budgetary restrictions together with some user behavior rules which are verified in an ad hoc implementation of the Yen’s k-shortest path algorithm [10].

The LSA solves as much instances of the optimization model as the number of lines under construction. Each instance builds only one line but determines the frequencies of all lines whose layouts (i.e. stretches and stations) are known (operating + already constructed + new constructed line). To this end, the constructed line in a given instance is transformed into a fictitious operating line for the next instance resolution. New lines are defined over the set of corridors built by the CGA. The instance optimization could be achieved with the aid of a solver based on either meta-heuristics or mathematical programming techniques. In this work, the latter approach is used under the premise that information about the quality of the solutions of the specific instances is available through the integrality gap metric of the Branch & Bound, and that the solving time is reasonable. In that manner, a quality standard for the instance solutions achieved by the LSA is guaranteed.

In each instance of the LSA, two goals are sought-after, minimize passenger riding time and minimize operator costs. To optimize these goals, goal programming [11] is used. This technique comes from multicriteria decision making and allows managing different and possibly conflicting objectives denominated goals in the context of an optimization problem. Current trends in urban planning emphasize the need for car-free cities [12,13], a concept that can only be realized by providing good level of service to passengers in public transport systems. Following this trend, this work assumes a strict preference for minimizing passenger riding time over minimizing operator costs. In the context of well-defined preferences, the use of Lexicographic Goal Programming is recommended [14], where goals are optimized sequentially according to their priority levels defined by the decision maker preferences. This technique has recently been used in some public transport works [15,16].

Two case studies based on the rapid transit networks of Santiago de Chile and Seville (Spain) are used for validation. Detailed solution reports are shown and discussed. The paper is organized as follows. Section 2 surveys the literature on TND and TNFS, highlighting the contributions of the proposed approach. Section 3 states the problem. Section 4 describes the modelling strategy. Section 5 introduces the formulation of the optimization model. Section 6 presents the solving strategy. Section 7 shows the computational tests. The paper finishes with some conclusions and directions for further research in Section 8.

2. Literature review

Bruno et al. [17] is one of the first to tackle the TND phase. Their work aims to maximize demand coverage in a public network. Laporte et al. [18] incorporate demand using an origin-destination trip matrix. The papers of Laporte et al. [19] and Hamacher et al. [20] address the stations location problem on a given alignment. García and Marín [21,22] study the mode interchange and parking network design problem using Bilevel Programming. They consider multimodal traffic allocation problems using combined modes at the lower level of the bilevel program. Laporte et al. [23] extend the previous models by incorporating into the station location problem the possibility to include the construction of several lines. The resulting model also considers budgetary constraints; however, line terminal nodes are fixed. Marín [24] overcomes this limitation.

References related to TNFS may be classified into the ones that consider the point of view of the operator and the ones that account for the point of view of the user. In the first group, Claessens [25] consider the minimization of the service costs. They formulate a model accounting for the selection of services, frequencies and train lengths, considering vehicle types. Cordeau et al. [26] also consider different types of train compositions. From the point of view of the user, Bussieck et al. [27] maximize the number of passengers without considering transfers. Scholl [28] minimizes the number of transfers by using passenger routes in the model and heuristics to generate feasible solutions. Schobel and Scholl [29] minimize travel time together with route selection from a predefined pool.

Works integrating TND and TNFS phases may be grouped according to the type of public transportation system to be planned. The vast majority of works deal with bus network design, whereas fewer works faces to the railway network planning. In the first group, Newell [30] studies the optimal geometries of bus routes depending on the demand trip distribution. The author proposes close formulas for computing desired frequencies depending on the choice of route geometries, and shows that the passenger cost function needed for route evaluation is not convex, therefore only local optimum can be found. Ceder & Wilson [1] formulate two sequential mathematical models for solving the bus network design problem. The first one minimizes the excess passenger travel time upon boarding, expressed as the deviation time from the shortest travel path, plus the transfer time (if any). The second model adds the passenger waiting time and vehicle operating and capital costs to the objective function. Baaj & Mahmassani [31] decompose the bus network design problem into three phases: a route generation step, in which routes and frequencies are constructed; a network analysis procedure, defining measures of effectiveness at the network-, route-, and stop-level; and a route improvement algorithm to improve the route design. These procedures are applied to different sets of weights, which examine the total travel time, total...

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
demand satisfied, and the required fleet size to operate the system. Cancela et al. [32] formulate a bilevel problem in which the upper level minimizes user costs subject to resource, station waiting and transfer constraints. In the lower level, the same objective is pursued but considering only constraints related to users.

Works dealing with advanced bus service design include the work of Furth & Wilson [33]. That work presents the problem of allocating buses to lines between time periods so as to maximize net social benefit. Wilson & Gonzales [34] focus on practice methods to identify bottleneck problems in the existing system.

Modifications on existing models are proposed to include multiple objectives. Results show significant differences when compared to standard models. Furth [35] devises a new service policy called deadheading where stops may be discontinuously served in a certain line service, i.e., odd line cycles may skip the stop but even cycles may serve that stop, or the other way round. The author also proposes asymmetrical services where stations may be skipped in some direction. Feillet et al. [36] propose a branch-and-cut-and-price method to efficiently construct multiple lines using express services where some intermediate stops can be skipped.

Deadheading is not allowed in such type of services. Leiva et al. [37] also determines the type of vehicle to be used and considers operation costs. Verbas & Mahmassani [38] integrate all the service policies.

In the group of railway planning approaches, Wan & Lo [39] formulate a mixed-integer programming problem that minimizes operation costs subject to resource constraints (fleet size, vehicle capacity and maximum stretch frequency). The model is validated on a small-sized network using CPLEX as solver. Borndörfer et al. [40] propose a multicommodity flow model where passengers are assigned according to their travel times, considering line infrastructure costs. Lines are dynamically built using a column generation algorithm that is applied to the city of Postdam, Germany. Canca et al. [41] present a non-linear mixed-integer programming problem, extending previous approaches by incorporating vehicle acquisition and setup costs in the context of competition between transportation modes. The model is solved with the aid of Conopt3 solver. Marin et al. [42] introduce the concept of “robustness”. To users, a robust network provides alternative routes in case of failures in their preferred routes. To operators, robustness is obtained by a better distribution of vehicles throughout the network, so that fewer vehicles are affected when a failure occurs. User robustness is tested on a high-speed network in Andalucia, a region in the South of Spain. Codina et al. [42,43] extend the previous work by incorporating competition between public transport and private car modes. The authors formulate a bilevel problem (BP) where the upper level designs the rapid transit network according to a given probability of failure that is evaluated by the lower level. This BP is applied to a test network.

A thorough review on transit planning works can be found in [2,44–48]. The review [47] puts emphasis on works integrating TND and TNFS phases from both modelling and solving strategy aspects. In this work, the following features previously studied in the literature are considered:

- Planning costs related to the construction of new infrastructure resources (new stations and new stretches) (Inf.), the acquisition of new vehicles (Veh.) and the frequency of vehicles on stretches (Freq.).
- An infrastructure budget (Inf. budget) limiting the construction of new infrastructure resources.
- Generation of two types of services: express services, where vehicles may skip some or all of the intermediate stations along a line; and local services, where vehicles halt at every intermediate station.
- Capacity limitations on the amount of passengers that a vehicle could carry, the number of vehicles going through a stretch, and the total number of available vehicles.
- Passenger times related to in-vehicle traveling (In-Veh.) and transfer between nearby stations.
- The imposition of a maximum waiting time at boarding stations to guarantee a minimum level of service to passengers.
- Penalization for not covering certain origin-destination demand pairs.

Table 1 highlights the features considered in the works mentioned in the context of railway systems. It is observed that none of these works consider all the features at the same time. Moreover, the infrastructure budget is only considered in a few works [4,41], and the express service design is tackled with for the first time.

This work also incorporates the following novel aspects to further improve the integration of TND and TNFS phases in the context of railway systems:

- The time a passenger waits inside the vehicle during a stop at an intermediate station. It is considered both in the objective function as a cost and in the computation of the line cycle time.
- The layover time at terminal stations. This is the time it takes the driver to go from one extreme of the train to the other in order to change driving direction.
- A pedestrian network representing not only passengers walking to nearby stations but also walking directly to destination.
- The ability to incorporate operating lines, whose layouts (stretches and stations) are already determined, but their operating frequencies can be changed.
- A planning budget limiting the acquisition of new vehicles.

Finally, and not least importantly, the problem of minimizing simultaneously the opposite objectives of passengers riding time and operator cost is addressed. As mentioned, the literature works either minimize solely one of these objectives or assess their relative importance in terms of weights. The latter approach is justified by the fact that the optimal solution for a multi-objective optimization problem represented with a weighted function is also a Pareto optimal solution. Nevertheless, achieving the optimal solution in real-world networks is not possible due to the complexity of integrating TND and TNFS phases. Therefore, the quality of the solution cannot easily be assessed. To overcome this limitation, the Lexicographic Goal Programming (LGP) technique is used. The LGP minimizes each objective hierarchically. First, the least passenger riding time is obtained within reasonable time while fulfilling an infrastructure budgetary constraint limiting the future investment to be carried out by the operator. Then, the operator cost is minimized while keeping the attained passenger riding time. This mechanism guarantees that the best effort for optimizing both objectives has been done attaching more importance to passengers goal.

3. Problem statement

The network design phase may start from scratch or may expand the railway rapid transit network, if there exist a set of lines already in operation. In this phase, the number of new lines together with their associated stations and stretches are determined. New lines are built from a pool of corridors with known stretches and terminal stations. A stretch is referred to as the section of a corridor between two consecutive stations. Several lines can be assigned to the same corridor, starting and finishing at the terminal stations of the corridor. However, these lines may halt at different intermediate stations. Fig. 1 depicts a corridor shared by three lines, composed of four stations. In the figure, stations 1 and 4 are terminal stations, whereas stations 2 and 3

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
Table 1

<table>
<thead>
<tr>
<th>Planning costs</th>
<th>Inf. Freq. Veh.</th>
<th>Capacity limitations</th>
<th>Passenger times</th>
<th>Level of service</th>
<th>Uncovered demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inf. Freq. Veh.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wan and Lo [39]</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
</tr>
<tr>
<td>Böndörfi et al. [40]</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td></td>
</tr>
<tr>
<td>Canca et al. [41]</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td></td>
</tr>
<tr>
<td>Marín et al. [42]</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td></td>
</tr>
<tr>
<td>Codina et al. [43]</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td></td>
</tr>
<tr>
<td>This work</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td>Check</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. A corridor shared by three lines, composed of four stations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. Service patterns provided by the lines assigned to the corridor shown in Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are intermediate stations. Additionally, the black line stops at all intermediate stations, whereas the blue line stops only at station 3. Finally, the green line does not serve any intermediate station.

The number of stretches and stations to be constructed is limited by an infrastructure budget. A limit on the maximum number of new lines to be constructed is also imposed.

The frequency setting phase determines the number of vehicles, the frequencies, and the service patterns to be provided by the set of lines (working plus new constructed lines). The term service refers to a complete line cycle where each station is traversed twice, once in each direction as shown in Fig. 2. The frequency of a line is then the number of services performed in a given time.

Vehicles may carry out two type of service patterns: local services and express services. In local services, vehicles halt at every station as in the black line on Figs. 1 and 2; whereas in express services, vehicles may skip some or all intermediate stations as in the blue and green lines on Figs. 1 and 2. The service time is made up of three time components: travel time, dwell time and layover time. Travel time refers to the time that a vehicle spends on traversing a stretch, whereas dwell time is associated with the time that a vehicle remains on a station. Finally, layover time refers to the time that it takes the driver to go from one extreme of the train to the other in order to change driving direction.

In a given line, travel and layover times are computed according to the corridor assigned to that line. However, service time is calculated according to the service pattern to be performed (as halting at intermediate stations increases, service time does). Fig. 2 shows the line cycle composition for the lines constructed in Fig. 1. The number of vehicles and services to be performed is limited by resource constraints (fleet size, vehicle and stretch capacities). A specified number of additional new vehicles may be purchased according to the planning budget in order to enlarge the fleet size.

The network design and the frequency setting phases are mutually dependent one from the other. For instance, if a vehicle halts at intermediate station i then station i must be built. Conversely, if no vehicle serves station i, then station i should not be constructed.

Railway rapid transit networks have to maximize passenger demand coverage. Passengers may transfer between lines by walking to nearby stations according to a maximum walking distance. Waiting time at stations is also limited and affects service level.

The problem described above is modeled and solved using mathematical programming techniques. The model decides on:

- Assignment of new lines to corridors.
- Construction of stations and stretches.
- Number of vehicles and frequencies provided by the set of new and existing lines.
- The type of service patterns (express or local) to be provided by the new lines.
- Passenger assignment to lines.

The following assumptions hold:

- There is a set of candidate corridors that is generated beforehand, using the corridor generation algorithm (described in Section 6.2).
- The fleet of vehicles is homogeneous, i.e., same capacity and average speed on stretches.
- Every line consists of one and only one service pattern (express or local). However, several lines may share the same corridor accounting for asymmetrical demand (i.e., in each direction the amount of passengers are different). Therefore, having different line frequencies.
- Average dwell time and layover time are known in advance.
- Average values for passenger boarding, alighting and transfer times are known in advance.

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
• Because the model is intended for systems where passengers determine in advance which lines need to choose, passenger assignment to lines is guided by the minimization of the
following time costs: in-vehicle traveling, in-vehicle waiting, maximum waiting time at boarding stations, alighting and transfer (described in Section 4.2). Waiting time prior to board
a vehicle is not explicitly modelled but limited by means of constraints (15), (19) stated in Section 5.5, and time costs at
boarding links are comprised of the boarding time and the
minimum amongst the maximum waiting time at stations and
a given time threshold characterizing the anticipation time of
passengers (typically few minutes).
• The number of trips for every origin-destination pair of sta-
tions (from the set of corridors) is known. This implies that if
a station is not constructed, the passengers that would have
departed from it will have to walk to a nearby station (using
transfer links).

4. Modelling strategy

The Network Design and the Frequency setting phases are
modeled using a graph, denominated Rapid Transit Graph (RTG).
Decisions regarding these phases are in direct correspondence
with the elements of this graph. As Rapid Transit Networks pro-
vide service to passengers, a directed graph called Passenger Flow
Network (PFN) is used to carry out passenger assignment to the
Rapid Transit System.

4.1. Rapid Transit Graph

The RTG holds the set of existing and candidate corridors. In
the graph, nodes represent stations, and links stretches. Stations
and stretches are either existing or candidate. Candidate stations
or stretches are resources that can be used by new lines if con-
structed. Existing stations and stretches are used by operating
lines, but can also be used by new lines. The RTG is undirected
because rapid transit vehicles are assumed to pass through sta-
tions twice (once in each direction) during a service on the line,
as explained in Section 3.

An illustrative example consisting of a RTG with five corridors
is shown in the following Figure. There exists one operating cor-
dor with constructed stations 1 to 4 and four candidate corridors:
6-1-5, 6-1-2, 5-1-2 and 8-3-7. In the figure, constructed station
1 can also be allocated to candidate corridors 6-1-5, 6-1-2
and 5-1-2, constructed stretch (1,2) to corridors 6-1-2, 5-1-2, and
constructed station 3 to corridor 8-3-7.

4.2. Passenger flow network

This graph is constructed as follows. For each station in the
RTG, an access node representing incoming and outgoing flows
of the station is used. Additionally, for each line likely to pass
through the station a pair of boarding and an alighting nodes are
used. The valid flows among these nodes are as shown in Fig. 4.

In this figure, passenger flow departing from a given access
node to any destination has to go through the set of connecting
links. Links can represent flow bifurcation and joint, as passengers
can exit the network or transfer between lines. The possible
links/flows are as follows. In-vehicle traveling links connect board-
ing nodes to alighting nodes of different and consecutive stations
(represented in this graph by their access nodes). This is independ-
ent on whether the vehicle actually halts at the station or not. If
the vehicle halts at the station, flows from the alighting node to
the access node, and flows from the access node to the boarding
node, may occur. Passengers flow that neither alight nor board at
the station and continue traveling is represented by the in-vehicle
waiting link. If the vehicle does not halt at the station, a skip-stop
link and flow is used. All the aforementioned flows are indexed by
the set of lines either new or in operation.

Since capacity resources and budget limitations are taken
into account, a feasible solution and even an optimal solution
to this problem may not provide full connectivity to passengers
using the RTN. As a consequence of that, transfer links and flows
connecting access nodes of nearby stations, and access nodes from
the origin station to the destination station, are used. These flows
are independent from the set of lines.

For lines in operation, at stations where vehicles halt, skip-stop
links are not considered. Conversely, at stations where vehicles
do not halt, only skip-stop links are taken into account. Fig. 5 shows
the PFN associated with station 3 of Fig. 3. This station is contained
in the corridors used by five different lines. Among these lines,
operating lines 1 and 2 serve station 3, whereas operating line 3
skips that station. New lines 6 and 7 are also considered and, for
these lines, skip-stop, in-vehicle waiting, alighting and boarding
links are used because their line layouts are not yet determined.

4.3. Integration of operator decisions and passenger assignment

The optimization model for the Network Design and Frequency
setting problem is built upon the elements of the RTG and PFN
graphs. Variables, parameters and sets are defined in direct cor-
respondence to graph elements. Decisions can be classified into
either operator decisions (layout, services and frequencies) or
passenger assignment (passenger flow distribution). Nevertheless,
both decision groups are interdependent, so the values of passen-
ger flow variables defined over the links in the PFN are constrained
by the line layouts and services established by variables associated
with the RTG, and vice-versa. This interdependency is illustrated
in the following Fig. 6.

In the figure, two possible scenarios for passenger assignment
are shown. Each scenario represents a different use of candidate
station i = 3 when line L6 is allocated to corridor c = 8 − 3 − 7.
Scenario 1 represents allowable flows when station L6 provides
service to passengers at station 3. Under this scenario, passenger
flow exchange at the access node of station 3 may occur; however,
skip-stop flow is forbidden. In scenario 2, line L6 does not provide
service to passengers at station 3. Therefore, flow exchange at
the access node is not allowed and incoming flow is directly
transferred to outgoing flow through skip-stop link.

5. Mathematical formulation

This section presents the formulation of the Network Design
and Frequency Setting model (NDFS) to be applied to railway
rapid transit networks. In this formulation, passenger flows are ex-
pressed for each origin of demand in order to reduce the problem
size. There are two objectives for this model, describing passenger
and operator interests. Goal programming will be later used to op-
timize the model, considering both objectives sequentially.

Definition of sets:

$L$ Set of lines.

$L_a$ Set of lines containing stretch $a \in ST$.

$L_i$ Set of lines associated with access node $i \in NW$ (i.e., the
line may provide service at station $i$ or pass by).

$L^D$ Set of operating lines.

$L^N$ Set of candidate new lines to be constructed in the Rapid
Transit Network.

$S$ Set of stations.

$S^N$ Set of candidate new stations to be allocated to new
lines $l \in L^N$.

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
Fig. 3. Rapid Transit Graph for a Rapid Transit System with one operating corridor and four candidate corridors.

Fig. 4. Passenger Flow Network representation for a candidate station on a new line.

Fig. 5. Passenger Flow Network representation for station 3 of Fig. 3.

Fig. 6. Allowable scenarios for passenger assignment on new line 6 at station 3 according to operator decisions.
$S^0(l)$ Set of stations to which operating line $l \in L^0$ provides service.  

$ST$ Set of stretches.  

$ST^l$ For an operating line $l \in L^0$, it is the set of stretches composing the line. For a new line $l \in L^N$, it is the set of stretches composing every candidate corridor to which the new line can be assigned.  

$ST^N$ Set of stretches that can be used by new lines $l \in L^N$.  

$C$ Set of corridors to which new lines $l \in L^N$ can be assigned.  

$c_a$ Set of corridors containing stretch $a \in ST^N$.  

$c_i$ Set of corridors containing station $i \in S^N$.  

$O$ Set of demand origins $(O \subseteq N_W)$.  

$D_p$ Set of destinations for demand originated at access node $p \in N_W$.  

$N$ Set of flow nodes.  

$N^l$ Set of flow nodes belonging to line $l \in L$.  

$N^a_l$ Set of alighting nodes belonging to line $l \in L$.  

$N^b_l$ Set of boarding nodes belonging to line $l \in L$.  

$N_W$ Set of access nodes.  

$A$ Set of flow links.  

$A_l^l$ Set of in-vehicle traveling links related to line $l \in L$.  

$A_l^i$ Set of in-vehicle waiting links related to line $l \in L$.  

$A_w$ Set of transfer links connecting nearby stations (according to a maximum allowable walking distance), and origin to destination stations.  

$A_w^l(i)$ Set of transfer links coming into access node $i \in N_W$.  

$A_w^l(i)$ Set of transfer links going out of access node $i \in N_W$.  

$A_w^i(l)$ Set of in-vehicle waiting links related to candidate new station $i \in S^N$.  

$A_w^i(l)$ Set of skip-stop links related to candidate new station $i \in S^N$.  

$A_w^i(l)$ Set of boarding links related to candidate new station $i \in S^N$.  

$A_w^a(l)$ Set of alighting links related to candidate new station $i \in S^N$.  

$A_w^{a,a}(l)$ Union set of in-vehicle waiting, boarding and alighting links related to candidate new station $i \in S^N$.  

$A_l^i(l)$ Set of boarding links from access node $i \in N_W$ to line $l \in L$.  

$A_l^a(l)$ Set of alighting links from line $l \in L$ to access node $i \in N_W$.  

$g_p$ Number of passengers per time unit originated at $p \in O$.  

$g_{pi}$ Number of passengers per time unit originated at $p \in O$ with destination in access node $i \in N_W$.  

$i_b$ Available budget for infrastructure resources.  

$k$ Constant related to the passenger arrival time distribution at stations.  

$q$ Vehicle capacity expressed as the maximum number of passengers that a vehicle can hold.  

$t_a$ Time cost associated with a link $a \in \{ST \cup A\}$.  

$t_l^t$ Layover time at any terminal station.  

$t_l^{d}$ Dwell time at station $i \in S$.  

$W_{max}$ Maximum allowable passenger waiting time at any station $i \in S$.  

**Definition of decision variables:**  

**Binary variables**  

$s^l_i$ Variable with value 1 when new line $l \in L^N$ provides service to passengers at station $i \in S^N$, value 0 otherwise.  

$x_a$ Variable with value 1 when new stretch $a \in ST^N$ is constructed, value 0 otherwise.  

$y_i$ Variable with value 1 when new station $i \in S^N$ is constructed, value 0 otherwise.  

$\delta_l^c$ Variable with value 1 when new line $l \in L^N$ is assigned to corridor $c \in C$, value 0 otherwise.  

**Integer variables**  

$b_l$ Number of vehicles working on line $l \in L$.  

$s^l$ Number of purchased vehicles.  

$z^l$ Number of services to be provided by line $l \in L$.  

**Continuous non-negative variables**  

$c^l$ Time that takes to complete one cycle (service) on the line $l \in L$.  

$f_l$ Number of vehicles per time unit working on line $l \in L$.  

$\Gamma_l$ Number of vehicles per time unit going through stretch $a \in ST^N$ on new line $l \in L^N$.  

$\lambda_l$ Number of vehicles per time unit halting at station $i \in S^N$ on new line $l \in L^N$.  

$u_p^a$ Portion of passengers originated at $p \in O$ traversing link $a \in A_l^l$. Value must be within $(0,1)$.  

$v_p^l$ Portion of passengers originated at $p \in O$ traversing link $a \in A_l^l$ of line $l \in L$. Value must be within $(0,1)$.  

**5.1. Passenger objective**  

$$z_{\text{pass}}(x,u) = \sum_{p \in O \setminus \lambda} \left( \sum_{a \in A_l^l} \sum_{s_i^l} t_{a,b} p_{a}^{l,s} + \sum_{s_i^l} t_{a,b} p_{a}^{l,s} + \sum_{s_i^l} t_{a,b} p_{a}^{l,s} + \sum_{s_i^l} t_{a,b} p_{a}^{l,s} \right)$$  

$$+ \sum_{p \in O \setminus \lambda} \sum_{s_i^l} t_{a,b} p_{a}^{l,s}$$  

(1)  

Passengers aim at minimizing their traveling times throughout the PFN network described in Subsection 4.2. These times include from left to right, the in-vehicle travel, the boarding, the alighting, the in-vehicle waiting, and the transfer times. Their associated costs per passenger are represented by parameter $t_{a,b}$. Passenger flows are expressed as portions of passengers originated at distinct demand origins with decision variables $u_p^a$ and $v_p^l$. The number of passengers in a given flow is obtained by multiplying this flow by the number of passengers originated at $p$, $g_p u_p^a$ and $g_p v_p^l$.  

**5.2. Operator objective**  

$$z_{\text{opt}}(x,y, \Delta b, \delta, s, f, b) = \sum_{i \in S^N} \sum_{a \in ST^N} c_i x_{i,a} + c_a \Delta b + c_l^f f_l$$  

$$+ \sum_{l \in L^0} \sum_{a \in ST^N} c_l^{l,f} f_l + c_b b_l$$  

(2)
Operator seeks to minimize costs related to infrastructure construction resources and exploitation costs. From left to right of (2), it is represented the construction costs of new station and stretches \((c_{y}|y, \text{ respectively})\), the acquisition of new vehicles \((\Delta b)\), the operation costs associated with working and new line frequencies \((c_{f{y}} f{l}_{c{y}}, \text{ respectively})\), and the setup costs of vehicles to lines \((c_{b} b{y})\). All these costs are expressed in monetary units per time unit.

### 5.3. Infrastructure budget constraint

\[
\sum_{i \in S^N} c_{y} y_{i} + \sum_{a \in ST^N} c_{x} x_{a} \leq ib
\]  

(3)

Constraint (3) sets the maximum number of infrastructure resources to be constructed according to the available infrastructure budget \((ib)\). The number of constructed resources is captured with the aid of decision variables \(y_{i}\) and \(x_{a}\).

### 5.4. Line generation constraints

\[
s_{l}^{i} \leq y_{i}, \quad \forall i \in S^N, l \in L^N
\]  

(4)

\[
s_{l}^{i} \leq \sum_{c \in C_{a}} \delta_{l}^{i}, \quad \forall i \in S^N, l \in L^N
\]  

(5)

\[
\delta_{l}^{i} \leq x_{a}, \quad \forall a \in ST^N, c \in C_{a}, l \in L^N
\]  

(6)

\[
\sum_{c \in C_{a}} \delta_{l}^{i} \leq 1, \quad \forall l \in L^N
\]  

(7)

Constraints (4) ensure that if a new line provides service to passengers at a new station, then that station must be constructed. Complementary, (5) verify that if a new line \(l\) provides service at a new station \(i\), then the line must be assigned to a corridor containing station \(i\). Analogously to (4), (6) ensure that if a new line \(l\) is assigned to corridor \(c\), then every stretch of the corridor must be constructed. Finally, (7) establish that a new line is only assigned to one corridor.

### 5.5. Line frequency setting constraints

The line frequency setting is based on the following relation:

\[
b{l^i} H \geq z{l^i} c{l^i}
\]  

(8)

where variables \(b{l^i}\) and \(z{l^i}\) represent the number of vehicles and the number of services to be provided by line \(l\), respectively, and \(c{l^i}\) is a function denoting the time that takes to complete one cycle (service) on the line \(l\). Finally, parameter \(H\) stands for the time horizon. In an optimal planning solution where setting vehicle costs are taken into account, it can be seen that \(b{l^i}\) is set to the minimum number of vehicles that are strictly needed to carry out the \(z{l^i}\) services on the line while not exceeding \(H\). This relation is implemented as follows. First, \(z{l^i}\) variable is replaced with the continuous variable \(f{l^i}\) defined as the number of services per time unit. Then, \(f{l^i}\) is accommodated into (8) by defining its relationship with \(z{l^i}\) as follows:

\[
z{l^i} = H f{l^i}
\]  

(9)

Consequently, Eq. (8) turns into the following:

\[
b{l^i} \geq f{l^i} c{l^i}
\]  

(10)

Now, function \(c{l^i}\) is substituted with its corresponding mathematical expression as follows:

\[
b{l^i} \geq f{l^i} \left( \sum_{a \in ST^N} i_{c} f{l^i} + 2 t{l^i} \right), \quad \forall l \in L^0
\]  

(11)

\[
b{l^i} \geq f{l^i} \left( \sum_{a \in ST^N} \sum_{c \in C_{a}} \delta_{l}^{i} \sum_{i \in L^N} \delta_{l}^{i} + 2 t{l^i} \right), \quad \forall l \in L^N
\]  

(12)

The line cycle consists of three time components: travel time, dwell time and layover time, as explained in Subsection 3. In lines under construction (12), travel and dwell times are not known beforehand because the corridor and the stations to which these lines are assigned and provide service, respectively, are decisions to be taken. Therefore, non-linearities given by products of \(f{l^i} \times \sum_{a \in ST^N} \delta_{l}^{i\text{ST}}\) and \(l^i \times \sum_{c \in C_{a}} \delta_{l}^{i\text{c}}\) arise. To overcome this problem, we employ Grover’s approach [49]. The author devised an exact method for linearising products involving binary and continuous non-negative variables. The method is used as follows. First, we define new variables \(f_{la} = f{l^i} \sum_{a \in ST^N} \delta_{l}^{i\text{ST}}\) and \(l_{la} = f{l^i} \sum_{c \in C_{a}} \delta_{l}^{i\text{c}}\), so (12) turns into the following:

\[
b{l^i} \geq \sum_{a \in ST^N} f_{la} f{l^i} + \sum_{c \in C_{a}} f{l^i} \delta_{l}^{i\text{c}} + 2 t{l^i} f{l^i}, \quad \forall l \in L^N
\]  

(13)

Second, we approach the relationships between new variables \(f_{la}\), \(f{l^i}\) and \(l_{la}\), \(l{l^i}\) as follows:

\[
f_{la} f{l^i} \leq f_{la} f{l^i} \leq f_{la} f{l^i}, \quad \forall a \in ST^N, l \in L^N
\]  

(14)

\[
f{l^i} f{l^i} (1 - s{l^i}) \leq f{l^i} f{l^i} \leq f{l^i} f{l^i} (1 - s{l^i}), \quad \forall i \in S^N, l \in L^N
\]  

(15)

\[
f_{la} \sum_{c \in C_{a}} \delta_{l}^{i\text{c}} \leq f_{la} \sum_{c \in C_{a}} \delta_{l}^{i\text{c}} \leq f_{la} \sum_{c \in C_{a}} \delta_{l}^{i\text{c}}, \quad \forall a \in ST^N, l \in L^N
\]  

(16)

\[
f{l^i} f{l^i} \left( 1 - \sum_{c \in C_{a}} \delta_{l}^{i\text{c}} \right) \leq f{l^i} f{l^i} \left( 1 - \sum_{c \in C_{a}} \delta_{l}^{i\text{c}} \right), \quad \forall a \in ST^N, l \in L^N
\]  

(17)

Constraints (14)-(15) state relationship \(f_{la} f{l^i} = f{l^i} \sum_{c \in C_{a}} \delta_{l}^{i\text{c}}\), whereas (16) and (17) represent relationship \(f_{la} f{l^i} = f{l^i} \sum_{c \in C_{a}} \delta_{l}^{i\text{c}}\). Upper bounds \(f_{la}\) and \(f{l^i}\) set the maximum number of vehicles per time unit halting at stations \(i \in S^N\) and going through new stretches \(a \in ST^N\), respectively. Analogously, lower bounds \(f_{la}\) and \(f{l^i}\) establish the minimum number of vehicles per time unit at each resource. The lower bound \(f{l^i}\) ensures a minimum level of service to passengers, setting a maximum passenger waiting time at any station \(i \in S^N\) prior to boarding a vehicle \((W_{max})\). The relationship between \(f{l^i}\) and \(W_{max}\) obeys the following formula:

\[
f{l^i} = \frac{k}{W_{max}}
\]  

(18)

where \(k\) is a parameter related to the passenger arrival distribution at stations. To maintain these minimum level of service to passengers in operating lines, the following set of constraints is also included:

\[
f{l^i} \geq \frac{k}{W_{max}}, \quad \forall l \in L^0
\]  

(19)

### 5.6. Planning capacity constraints

\[
\sum_{l \in L} b{l^i} \leq B + \Delta b
\]  

(20)
\( \Delta b \leq \Delta B \) \hfill (21)

\[
\sum_{l \in L} f^l_a \leq \tilde{f}_a, \quad \forall a \in ST
\] \hfill (22)

Constraints (20)–(22) fulfill some additional planning requirements. Constraint (20) ensures that the total number of vehicles set to lines does not exceed the current vehicle fleet size plus the number of acquired vehicles \((B + \Delta b)\). Additionally, (21) limits the number of purchased vehicles according to parameter \(\Delta B\). Finally, (22) limit the number of vehicles going through each stretch \(a \in ST\) according to its capacity \((\tilde{f}_a)\). Function \(f^l_a\) is defined as follows:

\[
f^l_a = \begin{cases} f^l & \text{if } l \in L^0 \\ f^l_\ell & \text{if } l \in L^N \end{cases}
\] \hfill (23)

5.7. Passenger flow balance constraints

\[
\sum_{l \in L} \left( \sum_{a \in A^L_{0}(i)} v^L_{a} - \sum_{a \in A^L_{1}(i)} v^L_{a} \right) + \sum_{a \in A^0} u^L_{a} - \sum_{a \in A^1} u^L_{a} - \sum_{a \in A^2} u^L_{a} = \begin{cases} 1 & \text{if } i = p \\ \frac{\sum_{i \in L} e^L_i}{e^L_p} & \text{if } i \neq p, i \in D_p, \forall i \in N_L, p \in O \\ 0 & \text{if } i \notin D_p \end{cases}
\] \hfill (24)

\[
\sum_{a \in A^L_{0}(i)} v^L_{a} - \sum_{a \in A^L_{1}(i)} v^L_{a} = 0, \quad \forall l \in L, i \in N_L, p \in O
\] \hfill (25)

Constraints (24) and (25) perform the passenger assignment in the PFN network as depicted in Fig. 4. Constraints (24) balance passenger flows connected to access nodes \(i \in N_L\), whereas (25) carry out the passenger flow balance at alighting and boarding nodes of each line \((i \in N^L_L \cup N^L_R)\).

5.8. Binding constraints

\[
\sum_{a \in A^L_{0}(i)} v^L_{a} \leq s^L_i, \quad \forall a \in L^0, i \in S^N, p \in O
\] \hfill (26)

\[
\sum_{a \in A^L_{1}(i)} v^L_{a} \leq s^L_i - s^L_{i'}, \quad \forall a \in L^N, i \in S^N, p \in O \quad \forall i, j \in S^N
\] \hfill (27)

Constraints (26) and (27) provide consistency between operator decisions and passenger assignment to new lines, as explained in Subsection 4.3. Constraints (26) ensure that passengers only alight and/or board lines at stations to which these lines provide service, whereas (27) prevent in-vehicle passengers from waiting at stations where lines do not provide service. If the line is assigned to a corridor not containing the station, then passenger flow balance neither occur at its alighting node nor at its boarding node.

5.9. Line capacity constraints

\[
\sum_{p \in P} v^L_{a} \leq q \tilde{f}^L_{at(a)}, \quad \forall l \in L, a \in A^L_v
\] \hfill (28)

Constraints (28) limit the maximum in-vehicle passenger flow on each stretch contained in the corridors to which the lines are assigned. That maximum is reached when equalizing the line capacity on that line, defined as the vehicle’s capacity \(q\) times the frequency of that line. Expression \(s(t)\) maps the in-vehicle piece \(a \in A^L_v\) to the corresponding stretch \(a \in S^L_v\). Finally, function \(\tilde{f}^L_{at(a)}\) is defined by (23).

5.10. Symmetry constraints

\[
\sum_{a \in S^N} \sum_{c \in C} \hat{f}^L_{ac} + \sum_{c \in C} \tilde{f}^L_{ac} + 2t \leq \sum_{a \in S^N} \sum_{c \in C} \hat{f}_ac + \sum_{c \in C} \tilde{f}^L_{ac} + 2t, \quad \forall i \in L^N \\{1\}
\] \hfill (29)

Constraint (29) breaks the line symmetry for those lines under construction, i.e., it gives an enumeration of the lines, so that line 1 must have a line cycle that is less than or equal to cycle of line 2 and so on. This constraint is not strictly necessary, but it speeds up the model resolution significantly.

5.11. Relaxing constraints

\[
y_i \leq \sum_{l \in L^N} s^L_i, \quad \forall i \in S^N
\] \hfill (30)

\[
x_a \leq \sum_{l \in L^N} s^L_i, \quad \forall a \in S^N
\] \hfill (31)

Constraints (30) and (31) prevent the emergence of fractional values of variables \(y_i\) and \(x_a\), when the integrality on them is relaxed. Again, these constraints are not strictly necessary, but they accelerate the model resolution significantly.

5.12. Model shortly formulation

The NDFS model may be expressed in terms of the above constraints and objective functions as follows:

\[
H(z_{\text{opt}}(X), z_{\text{opt}}(x)) \quad \{32\}
\]

s.t.: \( x \in X \)

where \(X\) represents the set of feasible solutions to the set of constraints given by (3)–(31), \(H(.)\) is a multi-criteria objective function that will be approached as described in Section 6.1.

6. Solving strategy

This section presents the solving strategy used for the NDFS model. Firstly, the application of the goal programming technique is described. Secondly, an algorithm to generate the pool of feasible corridors as input to the NDFS is introduced. Finally, a decomposition approach to efficiently solve the NDFS on real-world networks is presented.

6.1. Goal programming for the NDFS model

A goal programming problem takes the following general form:

\[
\min \ a = h(n, p) \quad \{33\}
\]

s.t.: \( f_q(x) + n_q - p_q = b_q, \quad q = 1 \ldots Q \)

\( x \in X \)

\( n_q \geq 0, \ p_q \geq 0, \quad q = 1 \ldots Q \)

where \(Q\) is the number of goals and \(f_q(x)\) is a function representing the value of goal \(q\) in terms of decision variable vector \(x\). Admissible values for \(x\) are contained in set \(X\). Finally, target values for each goal are represented by \(b_q\) and \(n_q\) and \(p_q\) are devitional variables representing the amount by which the left sides short of exceeds \(b_q\). These variables are minimized using an achievement function \(h(n, p)\).
There are several approaches to solve a goal programming problem. In this work, the Lexicographic variant of goal programming [14] is used, and is represented as follows:

\[
\text{Lex} \quad \min \ a = \left[ h_1(n_1, p_1), h_2(n_2, p_2), \ldots, h_l(n_l, p_l) \right] \quad (34)
\]

\[
f_q(x) + n_q - p_q = b_q, \quad q = 1 \ldots Q
\]

\[
x \in X
\]

\[
n_q \geq 0, \quad p_q \geq 0, \quad q = 1 \ldots Q
\]

The lexicographic variant assumes the existence of a hierarchy of goals, represented by an ordered set of priority levels \( L = \{1, 2, \ldots, |L|\} \), where a goal within priority level \( i \) is infinitely more important than a goal within level \( i+1 \). For each priority level, a function \( h_i \) over the values of its corresponding deviational variables \( n_i \) and \( p_i \) is defined. The form of each \( h_i \) is used to be linear and separable, so \( h_i \) represents the priority level. The meaning of lexicographic minimization of an achievement function \( a \) (defined over the \( |L| \) priority levels) is that the minimization of \( h_i \) is infinitely more important than the minimization of \( h_j \) for a lower level \( j \). As a consequence of this, first only \( h_1 \) is minimized, then the best value for \( h_1 \) is maintained during the minimization of \( h_2 \), and so forth. A thorough description of this technique can be found in [50].

The Lexicographic Goal programming (LGP) variant is applied to the NDFS model (32) as follows. Let \( f_1 \) represent the goal of minimizing passenger riding time and \( f_2 \) the goal of minimizing operator costs, so that \( f_1 \) is strictly preferred over \( f_2 \). Let also \( z_{\text{pax}}(x) \) and \( z_{\text{op}}(x) \) be the functions denoting the value of goals \( f_1 \) and \( f_2 \), respectively, as defined in (1) and (2) for a given decision variable vector \( x \) belonging to its feasible set \( X \) (i.e., fulfilling every constraint of the model). In this work, the goal targets (not necessarily attainable) are \( b_1 = b_2 = 0 \), then the first goal programming stage is:

\[
\min n_1 + p_1 \quad (35)
\]

\[
\text{s.t.}: \quad z_{\text{pax}}(x) + n_1 - p_1 = 0 \quad x \in X
\]

As \( z_{\text{pax}}(x) > 0 \), \( \forall x \in X \) (i.e., it is impossible that passenger riding time fall under a zero value), then deviational variable \( n_1 = 0 \) and \( z_{\text{pax}}(x) = p_1 \), \( \forall x \in X \). So, stage 1 can be rewritten as:

\[
\min z_{\text{pax}}(x) \quad (36)
\]

\[
\text{s.t.}: \quad x \in X
\]

Under a similar reasoning, stage 2 can be formulated as follows:

\[
\min z_{\text{op}}(x) \quad (37)
\]

\[
\text{s.t.}: \quad z_{\text{pax}}(x) = z^*_{\text{pax}} \quad x \in X
\]

where \( z^*_{\text{pax}} \) is the optimal or best value found when optimizing the passenger riding time, and it is set as a constraint. Therefore, once stage 2 is solved, both goals have been optimized under the stated preference structure.

6.2. The corridor generation algorithm

The pool of candidate corridors to be assigned to new lines during the optimization process is previously determined by a procedure called the Corridor Generation Algorithm (CGA). This procedure implements a constrained version of the Yen’s k-shortest path algorithm [10] over an undirected graph consisting of every station and stretch (candidate or already constructed). The CGA is applied to every relevant pair of stations \( i, j \in S \), i.e., to those stations with significant attracted/originated demand, resulting in \( K \) shortest paths that later will be considered as corridors. During the application of this algorithm, a generated path will be considered as a corridor only if it satisfies the following constraints:

**CGA constraint 1.** Let \( y_{ij}^k \) be the time associated with the shortest path between a pair of nodes \( i, j \in S \), then for every pair of nodes contained in a given path \( k \), the following rule is checked:

\[
y_{ij}^k(1 + \Delta_L) \leq y_{ij}^k \leq y_{ij}^k(1 + \Delta_U)
\]

where \( \Delta_L \) and \( \Delta_U \) are nonnegative parameters denoting the minimum and maximum deviation times from the shortest path between nodes \( i \) and \( j \), and \( y_{ij}^k \) is the time associated with the subpath in \( k \) connecting them. This constraint is taken from [52] and ensures that every subpath contained in a corridor does not overlap/deviate too much with/from the shortest path.

**CGA constraint 2.** The sum of the construction cost of links (stretches) and nodes (stations) in a path representing a candidate corridor must not exceed the infrastructure budget \( b \) as in constraint (3).

**CGA constraint 3.** The attracted demand in a path \( k > 1 \) must be greater than the attracted demand in the shortest path \( k = 1 \). If the path \( k \) is represented by the ordered sequence of nodes \((i_1, i_2, \ldots, i_l)\), this constraint can be expressed as follows:

\[
\sum_{j=1}^{n} \sum_{j=1}^{n} g^k_j > \sum_{j=1}^{n} g^1_j
\]

where \( g^k_j \) is the demand originated at candidate station \( i \).

Table 2 shows the pseudo-code of the CGA for a given pair of stations \( i, j \in S \). It employs: the in-vehicle travel times matrix \( T \); list \( B \), which contains all the computed feasible paths; and the sublist \( Q \subset B \), which contains all the computed feasible not chosen as some k-shortest paths in \( P_i^k \). The subroutine begins by initializing list \( Q \) to null and list \( B \) to the shortest path \( P_1 \). Then, the iterative part of the algorithm starts. Every k-iteration seeks all the existing feasible new paths, such that they contain a common subpath \( (P_k - i) \) coming from the k − 1 shortest path \( P^{k-1} \), which itself starts at the source node \( P_1 \) and ends at an intermediate node \( P^{k-1} \). However, they differ from a detour subpath which starts at \( P^{k-1} \) and ends at terminal node \( P^{k-1} \). Feasibility is checked by verifying if every new path satisfies the aforementioned planning and construction constraints, as well as the user input rules. To compute these new subpaths, the method works with a copy of the in-vehicle travel times matrix \( T \). This copy may be modified according to the position of the intermediate node \( P_i^{k-1} \) (under process) in path \( P^{k-1} \) and the links held in the common subpath \( P^{k-1} - i \). This modification is carried out in two steps:

1. The links adjacent to \( P_i^{k-1} \) are eliminated. These links are contained in the paths of list \( B \). Their initial subpath from the first node \( B_i^j \) to the intermediate node \( B_j^i \) overlaps with the initial subpath of k − 1, from its first node \( P_i^{k-1} \) to the intermediate \( P_j^{k-1} \). (81)

2. The links adjacent to nodes in the initial subpath of path \( k − 1 \) are eliminated, from its first node \( P_i^{k-1} \) to the antecessor node of the intermediate node \( P_i^{k-1} \). (82)

The second step mentioned above is not specified in the original Yen’s algorithm [10], but it is required to deal with undirected graphs to prevent more than one visit to any node already contained in the initial subpath \( P^{k-1} \).
Having updated $T'$ properly, the shortest path from $p_{k-1}^{i-1}$ to $p_{k}^{i-1}$ is computed by using Dijkstra's algorithm [51]. If a new subpath ($Z$) linking the intermediate node $p_{k}^{i-1}$ to the final node $p_{k}^{i-1}$ is found, the algorithm verifies whether $Z$ satisfies rule (38). Firstly, the right hand side inequality of (38) is checked by comparing the travel time cost of detour $Z$ ($y_{ln}^Z$) with the travel time cost of the k−1 shortest path ($y_{ln}^{k-1}$) times 1 + $\Delta U$. If $y_{ln}^Z$ is not from above, then a new path $R$ is built by appending $Z$ to the initial subpath $p_{k}^{i-1}$ and the left hand side inequality of (38) is verified as follows. For every pair of nodes $p$, $q$, such that $p$ is held in the common subpath ($p \in p_{k}^{i-1}$) and $q$ is contained in the detour subpath ($q \in Z_{2-3}$), it is compared the travel time cost from $p$ to $q$ in the new subpath ($y_{ln}^{Z}$) with the travel time cost of the shortest path from $p$ to $q$, $y_{ln}^{k-1}$ times (1 + $\Delta U$). If $y_{ln}^Z$ is from below, the inequality is satisfied. If any of these two conditions are satisfied, path $R$ is rejected and a new k-iteration is evaluated. Finally, $R$ is verified for whether or not it satisfies constraints (3) and (39). This is done by means of the mapping functions $c(R)$ and $g(R)$, respectively. If so, it is appended to lists $Q$ and $B$. Otherwise, it is also rejected.

To complete the kth-iteration, the shortest travel time is chosen from the list of candidate paths $Q$ and it is then set to the k-shortest feasible path. Finally, this chosen path is deleted from $Q$. The process is repeated at most K − 1 times, providing that the list $Q$ is not empty when it tries to set a k-shortest path, with $k < K$. If so, the algorithm is finished earlier.

**6.3. The Line Splitting Algorithm**

Solving the LGP formulation of NDFS model for multiple lines under construction turns out to be computationally inefficient, preventing the application of the model on real-world networks. To overcome this limitation a decomposition strategy called Line Splitting Algorithm (LSA) is introduced.

The LSA solves as much instances of the LGP-NDFS model given by (36) and (37) as the number of lines under construction ($|L^q|$). Therefore, in each instance, only one line is constructed.

Having solved the last instance, a global feasible solution to the original LGP-NDFS is obtained. Two variants of the algorithm are developed. One in which all the demand is assigned at every instance, and another where the amount of the demand is assigned incrementally according to the number of lines already considered. For instance, on iteration $n$, $\frac{n}{|L^q|}$ of the demand is considered.

Fig. 7 shows an application of the LSA for a case with 3 operating lines (labeled as L1, L2 and L3) and 4 candidate new lines (labeled as L4, L5, L6 and L7). The LSA solves 4 LGP-NDFS instances in order to construct all new lines.

A detailed description of the matheuristic is shown in Table 3. The LSA takes as inputs all data sets and parameters that conform to the global graph (G), the OD-demand vector ($\mathbf{g}$). An additional parameter $\alpha_g$ is included to establish how the demand is assigned.

The algorithm can be split into two phases. The first phase consists of initialising the data sets and parameters related to the network layout, which are modified in each iteration of the algorithm, and initialising the incremental load portion $\Delta g_p$, which is used later to update the passenger demand.

The second phase is an iterative process in which the instances of the LGP-NDFS model are solved. Each instance consists of a reduced mathematical model where only a new line is constructed.
according to the nodes and stretches given by sets $S$ and $ST$. The rest of the lines included in the resolution are treated as operating lines, having arisen from the heuristic input or from previous solved instances (fictional operating lines).

Having solved each instance, a backup of decision variables corresponding to the constructed line layout is made, as well as an appropriate update of its associated infrastructure data. In this update, the infrastructure budget is reduced according to the new stretches and stations that have been constructed, determined by variables $x_i$ and $y_j$. Right after, the construction costs associated with used resources (i.e., $x_i = 1$ and $y_j = 1$) are set to 0, so that these costs are not considered for subsequent resolutions of LGP-NDFS instances, although these constructed stretches and stations can also be allocated to further new lines. In the Rapid Transit Network considered in Fig. 7, 4 new candidate lines labeled as L4, L5, L6, and L7 can be constructed. Suppose that in the first instance resolution a line containing candidate corridor 6-1-2 shown in Fig. 3 is built. Moreover, assigned vehicles only halt at stations 6 and 2. Then, for the following instance resolutions, the construction costs associated with stretches 6-1-2, and stations 6 and 2, are set to 0 and the infrastructure budget is reduced according to the costs of these resources.

Finally, the data sets related to the stations and stretches of the constructed line ($S^f(l)$ and $ST^f$, respectively) are created, and line sets ($L^f$ and $L^{\theta}$) are updated in order to treat the constructed line as an operating line for the subsequent instance resolutions. In each instance resolution, the number of vehicles and the services performed at every line are recomputed so that the previous vehicles and services assigned to it are not considered.

### 7. Numerical experiments

This section reports the experiments carried out on the LGP-NDFS model. First, a comparison between the application of the LSA and the Branch & Bound algorithm of CPLEX to the LGP-NDFS model is performed on a test network. This comparison allows analyzing the quality of the solutions obtained by the LSA. The LSA proves to be useful, therefore its application to two real-world networks is then reported.

The LGP-NDFS is solved calling CPLEX 12.4.0 in a R5500 workstation with processor Intel(R) Xeon(R) CPU E5645 2.40 GHz, and 48 Gbytes of RAM. The CGA and LSA algorithms are implemented in MATLAB, which calls CPLEX in order to solve the successive instances of the LGP-NDFS.

In the following subsections, the description of the test network and the two case studies is presented, together with the results.

#### 7.1. Test network

Fig. 8 illustrates the test network used for comparison purposes between solving the LGP-NDFS model with/without the LSA. This network has been previously used in [4,24], among other works. In the figure, there is a four component vector $(c^l_0, d^c_t, c^l_1, f_0)$ attached to each stretch. The meaning of each vector component, from left to right, is as follows: the construction cost, the in-vehicle traveling distance, the operation cost, and the capacity (maximum number of vehicles per hour). Additionally, there is a number next to the nodes denoting the construction cost of the station. Finally, the frequency lower bounds attached to each resource are set to 0.

Moving on to demand issues, the amount of demand is 373 thousands passengers/h, and is divided into each pair of access nodes as follows:

$$g = \begin{bmatrix} 4 & 3 & 9 & 7 & 5 & 4 & 2 & 2 & 2 \\ 10 & 7 & 10 & 8 & 3 & 3 & 4 & 4 \\ 7 & 3 & 4 & 8 & 6 & 7 & 6 & 6 \\ 5 & 5 & 3 & 3 & 7 & 4 & 6 & 6 \\ 9 & 1 & 8 & 8 & 5 & 4 & 10 & 10 \\ 3 & 2 & 8 & 7 & 5 & 6 & 5 & 5 \\ 2 & 3 & 4 & 6 & 7 & 9 & 6 & 7 \\ 2 & 3 & 4 & 6 & 4 & 7 & 5 & 7 \end{bmatrix}$$

For passenger walking times, all combinations between station pairs have been considered. These times are computed as the minimum road distance divided by an average passenger speed of 4.5 km/h, resulting in the following walking time matrix:

$$T_w = \begin{bmatrix} 1.6 & 0.8 & 2 & 1.6 & 2.5 & 4 & 3.6 & 4.6 \\ 2 & 0.9 & 1.2 & 1.5 & 2.5 & 3 & 2.9 & 2.9 \\ 1.5 & 1.4 & 1.3 & 0.9 & 2 & 3.3 & 2.9 & 2.9 \\ 1.9 & 2 & 1.9 & 1 & 1.5 & 2 & 2.3 & 3.8 & 4.1 \\ 3 & 1.5 & 2 & 2 & 1.5 & 3 & 2.3 & 2.3 & 2.3 \\ 3.9 & 3.9 & 3.9 & 2 & 3 & 2.5 & 2.5 & 2.5 & 2.5 \\ 5 & 3.5 & 4 & 4 & 2 & 3 & 2.5 & 2.5 & 2.5 \\ 4.6 & 4.5 & 4.5 & 3.5 & 3 & 2.5 & 2.5 & 2.5 & 2.5 \end{bmatrix}$$

The average transfer and in-vehicle waiting times per passenger have been set to 4 min, whereas the average exiting time per
passenger is 2 min. Regarding operator issues, the infrastructure budget amounts to 40,000 mu/h and up to 4 new vehicles can be acquired at a unitary acquisition cost of 20 mu/h. The total vehicle’s capacity amounts to 550 passengers and its average working speed (without considering dwell time at stations) rises 40 km/h.

The available vehicle fleet is set to 8 vehicles and each one has a setting cost of 80 mu/h (monetary units per hour). Their average dwell time per station is considered to be 1 min, but layover times are neglected. Finally, no operating lines have been considered.

7.2. Seville underground

Fig. 9 illustrates the Seville underground, a medium-sized network made up of 24 nodes, 264 links and 552 OD-demand pairs, with a total of 24,446 trips/h in a working day. All these nodes and links represent new network infrastructure resources as there are no operating lines.

Resource parameters have been set with the aid of [53] as follows. Stretches have an average length of 350 m, with minimum and maximum values of 200 and 600 m, respectively. The construction cost per km of stretch is 189 millions of euros, and must be amortized in 25 years. This is the time in which the infrastructure resources depreciate. This amortization is taken into account using the net present value formula with no initial investment and a discount rate of 14%. Having applied this formula, an average construction cost of 365 €/h and minimum and maximum values of 60 and 583 €/h are obtained. As for stations, the construction costs were obtained estimating the platform area according to the attracted demand, and using an average cost for building 1 m² of platform. The resulting construction costs are between 176 and 374 €/h, with an average value of 244 €/h. Finally, the number of resources that can be built is limited to an infrastructure budget of 15,000 €/h.

Regarding operation features, stretches are homogeneous with a capacity of 12 veh/h and an operation cost of 20 €/h. The available planning budget allows buying up to 20 vehicles. Each vehicle has a setting cost of 20 €/h and can hold a total of 275 passengers. Its average working speed is 30 km/h, going through the stretches, but this speed is reduced with an average dwell time per station of 1 min and an average layover time at terminal stations of 2 min. Additionally, new vehicles can be acquired at a cost of 80 €/h each one.

The passenger assignment settings are as follows. Passengers have been estimated to walk up to 500 m, with an average speed of 4.5 km/h. The walking distance between stations has been considered the same as the stretch length for the sake of simplification. The average transfer and in-vehicle waiting times per passenger have been set to 4 min, whereas the average exiting time per passenger is 2 min.

7.3. Santiago de Chile underground

Fig. 10 illustrates Santiago de Chile underground, a more realistic network containing 146 nodes, 520 links and 21,316 OD-demand pairs, with a total of 3195 trips/min. At present, there are five working lines, labeled as L1, L2, L4, L4A and L5. Their infrastructure resources amount to 100 stations, from which 8 are transfer stations (stations where passengers can move from one line to another), and 206 stretches.

The new infrastructure resources for the network expansion amounts to 53 stations and 318 stretches, and are distributed in three subgraphs in order to reduce the number of transfers for those OD-pairs with origin and/or destination in some extreme or near-extreme nodes of lines L2, L4, L4A and L5. In this way, the subgraph labeled as SG3 allows linking the L5 terminal node, Plaza de Maipú (075), to the L2 terminal node, La Cisterna (048); whereas the subgraph SG1 connects the L2 terminal node, Vespucio Norte (028), to the L1 terminal node, San Pablo (082). Finally, the subgraph SG2 links the L1 terminal node, San Pablo, to the L2 terminal node, Vespucio Norte.

Resource parameters have also been set with the aid of [53] as follows. Stretches have an average length of 1.6 km and minimum and maximum values of 0.5 and 8.1 km, respectively. The cost of
1 km of stretch is 18.5 €/min, using the net present value formula with an amortization period of 25 years, no initial investment, and a discount rate of 12%. Thus, the construction costs are between 11.5 and 69 €/min, with an average value of 30 €/h. As for stations, the construction costs were not obtained by estimating the cost of the platform area as the actual attracted demand was not known beforehand. Instead, an estimated heterogeneous value of 3.25 €/min has been set. Finally, the number of resources that can be built is limited to an infrastructure budget of 9657 €/min.

Regarding operation features, stretches are homogeneous, so they have different capacities and operation costs. Their values comprised 0.1-0.5 veh/min and 3–50 €/min, with average values of 0.3 veh/min and 10 €/min, respectively. The current working fleet holds 188 vehicles, but can be extended up to 94 vehicles. Setting costs have been neglected, but newly purchased vehicles have an acquisition cost of 0.5 €/min. Each vehicle can hold a total amount of 1134 passengers, and its average working speed is 60 km/h. Finally, the average dwell time per station and average layover time at terminal stations have the same values as those of the Seville network.

The passenger assignment settings are also the same as for the Seville network.

7.4. Results

The reported results on the described networks show the impact of the number of candidate lines to be constructed and the maximum allowable passenger waiting time on the computational performance of the LGP-NDFS model. Prior to solve this model, the CGA is applied to determine the pool of candidate corridors as follows. In the test network, the CGA is applied to every pair of stations with $K = 17$, $\Delta_L = 0.55$ and $\Delta_U = 4.8$, resulting in 295 corridors. The parameters were calibrated such that the optimal line traces obtained by previously solving an optimization model incorporating the line routing were encountered within the pool of generated corridors [55].

As for the real-world networks, parameters $\Delta_L$ and $\Delta_U$ are set to 0.4 and 0.5, respectively, based on the practical experimentation undertaken by Schnabel and Löhse [54]. This is the only reference we have found related to the setting of the user's behavioral rules (38). In a real-world application of these rules, a specific survey on the future user population of the network should be conducted in order to provide reliable values for $\Delta_L$ and $\Delta_U$ parameters. Nevertheless, this matter is out of the scope of this work and the authors believe it does not diminish its applicability.
Finally, parameter $K$ is set as follows. In the Seville network, the
CGA was applied to a subset of station pairs where the levels of
generated/absorbed demand were significant. To be more precise,
those stations generating/attracting low levels of demand were
discarded, so that only the pairs from stations with high demand
generation to stations with high demand attraction were consid-
ered, resulting in 160 station pairs. Using this subset and verifying
all CGA-constraints at the same time, up to 478 feasible corridors
were generated setting $K \geq 3$. In Santiago de Chile network, the
CGA was only applied to the station pairs connecting the extreme
nodes of each subgraph $SG_i$ as the aim is to reduce the number
of transfers for those OD-pairs with origin and/or destination
in these stations. Up to 100 feasible corridors were determined,
setting $K \geq 35$.

Table 4 reports the problem size according to the network and
the number of lines under construction. The table includes the
number of discrete and continuous non-negative variables, the
cnumber of constraints, and the number of non-zero coefficients in
the constraint matrix. In the constraints column, the “+1” denotes
the extra constraint added to the LGP-NFDS when solving the
stage 2 (i.e., to maintain the optimal or best solution found for
the passenger goal).

The results show that the size of the problems increase almost
linearly to the number of lines under construction.

Next Table 5 shows a summary of the results obtained for the
experiments performed on the test network. In these experiments,
the number of candidate lines to be constructed ($|L^H|$) has been
varied from 2 to 4, and the maximum allowable waiting times
($W_{max}$) from 3 to 12 min. For each experiment, it is first reported
the optimum goal function value ($z_{GOM}$) obtained by solving
the LGP-NDFS Model with the Branch & Bound algorithm of CPLEX
(B & B), as well as the best goal function value provided by the LSA
variants. Then, the difference between $z_{GOM}$ values found by the
LSA variants and B & B resolution is shown, together with the total
time in seconds $T_{TOTAL}(s)$ for each solving method. LSA variants
are identified by labels “LSA1” and “LSA2” referring to LSA with incre-
mental demand load and LSA with fixed demand load, respectively.
Finally, goals are represented by labels “pax” and “op” standing for
the passenger riding time and operator costs, respectively.

The results show that as the values of $|L^H|$ and $W_{max}$ increase,
do CPU times for the resolution using B & B. Specially for $W_{max} = 12$ min where each increase of one line under construction
entails consuming one extra order of magnitude in CPU time.
However, CPU times of the LSA variants are always within a few
seconds. As for the solution quality, the LSA variants report either
the optimal value of $z_{GOM}$ or a near-optimal one. In the latter
case, reported solutions have a gap difference of at most 5% in $z_{GOM}$
values as shown in the experiments with $|L^H| = 2$ and $W_{max} = 6$, and
$|L^H| = 2$ and $W_{max} = 12$. Additionally, it is observed in experi-
ments with $|L^H| = 3$ and $W_{max} = 12$ that gaps for $z_{GOM}$ have low
and negative values as reported values for $z_{GOM}$ by the LSA variants
are lower than those found by using B & B. This issue may occur
when passenger riding times computed by the LSA variants are
higher than those obtained by using B & B. Comparing the LSA
variants, it is not very clear to determine which variant performs
better as their average gaps and CPU times are rather similar.

Moving on to the analysis of the LGP-NDFS model on the real-
world networks, Tables 6 and 7 report the performance results of
B & B and the LSA variants on instances varying $|L^H|$ from 2 to 3,
and $W_{max}$ from 3 to 12 min. The negativity in the relative gap
measure indicates that the B & B gives worse solutions compared
to the ones provided by the LSA variants when B & B does not
reach optimality within the time limit of 1 day. Furthermore,
operator goal is neither reported in instances with $|L^H| = 3$ in
Table 6 nor in Table 7 as the B & B spent all solving time on opti-
mizing the passenger goal. So, in those instances, only passenger
goal values are compared, and the solving times reported for the
LSA corresponds only to the time spent on optimizing this goal.

The results show that the LSA variants always attain the same
or slightly better values for the passenger goal, whereas
the solutions provided for operator goal, when applicable, are almost

---

**Table 4**

Size of the mixed-integer linear goal-programming problems solved in the reported experiments.

| Network | $|L^H|$ | Discrete variables | Continuous variables | Constraints | Non-zero elements |
|---------|--------|--------------------|----------------------|-------------|------------------|
| Test    | 2      | 560                | 4231                 | 2149 + 1    | 23310            |
|         | 3      | 840                | 6004                 | 3164 + 1    | 33987            |
|         | 4      | 1120               | 7777                 | 4176 + 1    | 44658            |
| Seville | 1      | 518                | 7811                 | 10940 + 1   | 281394           |
|         | 2      | 10036              | 142272               | 20075 + 1   | 521119           |
|         | 3      | 1554               | 208430               | 29723 + 1   | 762360           |
| Santiago | 1     | 152               | 576824               | 162887 + 1  | 2052033          |
| de Chile | 2     | 305               | 1036214              | 238238 + 1  | 3679077          |
|         | 3      | 440               | 1304248              | 303958 + 1  | 4678713          |

---

**Table 5**

Comparison between solving the full LGP-NDFS model directly and applying the Line Splitting Algorithm on the test network.

| $|L^H|$ | $W_{max}$ | Goal | $z_{GOM}$ | GAP$_{GOM}$ | $T_{TOTAL}(s)$ |
|--------|-----------|------|-----------|-------------|----------------|
|        | B & B     | LSA1 | LSA2      |             |               |
|        | B & B     | LSA1 | LSA2      |             |               |
| 3      | 3 min     | pax  | 185,409   | 185,409     | 185,409       | 0% | 0% | 1 | 2 | 2 |
|        |           | op   | 3898      | 3898        | 3898          | 0% | 0% |   |   |   |
| 2      | 6 min     | pax  | 179,305   | 179,314     | 179,305       | <1% | <1% | 4 | 2 | 2 |
|        |           | op   | 6724      | 7085        | 6865          | 5% | 2% |   |   |   |
| 12     | 6 min     | pax  | 178,700   | 179,191     | 179,186       | <1% | <1% | 21 | 13 | 7 |
|        |           | op   | 8517      | 9626        | 9636          | 5% | 2% |   |   |   |
| 3      | 6 min     | pax  | 185,409   | 185,409     | 185,409       | 0% | 0% | 1 | 2 | 2 |
|        |           | op   | 3898      | 3898        | 3898          | 0% | 0% |   |   |   |
| 12     | 6 min     | pax  | 173,252   | 173,581     | 173,577       | <1% | <1% | 965 | 5 | 5 |
|        |           | op   | 10,252    | 12,233      | 10,237        | <1% | <1% |   |   |   |
| 3      | 6 min     | pax  | 185,409   | 185,409     | 185,409       | 0% | 0% | 1 | 3 | 3 |
|        |           | op   | 3898      | 3898        | 3898          | 0% | 0% |   |   |   |
| 4      | 6 min     | pax  | 176,199   | 176,199     | 176,199       | 0% | 0% | 2 | 2 | 3 |
|        |           | op   | 7855      | 7996        | 8056          | 0% | 0% |   |   |   |
| 12     | 6 min     | pax  | 169,468   | 170,234     | 169,468       | <1% | <1% | 6921 | 6 | 6 |
|        |           | op   | 11,371    | 11,312      | 11,432        | <1% | <1% |   |   |   |
the same. Furthermore, in Seville network, solving times of LSA variants are one order of magnitude less than the ones of B & B; whereas, in Santiago network, solving times of LSA variants are far away from the time limit. The results show that the cumulative time to reach the best solution is one order of magnitude less than the total time (as can be observed in the summary rows). In Seville network, total time is around 1 h but only 10 min of cumulative time is needed to found the solution, approximately. As for Santiago de Chile network, total time is around 1 day but only 2 h is needed to reach the solution. The range of the times to reach the optimal solutions in each LSA iteration are rather similar in Santiago de Chile network. However, in Seville network, they are not so similar, specially when applying the LSA with incremental demand load. No matter the LSA variant used, the largest variation in the time range occurs in the first iteration of the LSA variant. Additionally, it is observed that computational times are not really affected by

### Table 6
Comparison between solving the full LGP-NDFS model directly and applying the Line Splitting Algorithm on Seville underground.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>(W_{\text{min}})</th>
<th>Goal</th>
<th>(z_{\text{GOL}})</th>
<th>(G_{\text{P\text{GOL}}})</th>
<th>(T_{\text{TOTAL}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10^3)</td>
<td>B &amp; B</td>
<td>LSA1</td>
<td>LSA2</td>
<td>B &amp; B</td>
<td>LSA1</td>
</tr>
<tr>
<td>3 min</td>
<td>5944</td>
<td>5944</td>
<td>5944</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>6 min</td>
<td>5944</td>
<td>5944</td>
<td>5944</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>12 min</td>
<td>5944</td>
<td>5944</td>
<td>5944</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>3 min</td>
<td>5480</td>
<td>5477</td>
<td>5477</td>
<td>–3%</td>
<td>–3%</td>
</tr>
<tr>
<td>6 min</td>
<td>5639</td>
<td>5477</td>
<td>5477</td>
<td>–3%</td>
<td>–3%</td>
</tr>
</tbody>
</table>

### Table 7
Comparison between optimizing the passengers goal with B & B of CPLEX and with the Line Splitting Algorithm on Santiago underground.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>(W_{\text{min}})</th>
<th>(z_{\text{opt}})</th>
<th>(G_{\text{P\text{opt}}})</th>
<th>(T_{\text{TOTAL}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10^3)</td>
<td>B &amp; B</td>
<td>LSA1</td>
<td>LSA2</td>
<td>B &amp; B</td>
</tr>
<tr>
<td>3 min</td>
<td>456,822</td>
<td>456,822</td>
<td>456,822</td>
<td>0%</td>
</tr>
<tr>
<td>6 min</td>
<td>245,218</td>
<td>237,255</td>
<td>237,255</td>
<td>–3%</td>
</tr>
<tr>
<td>12 min</td>
<td>250,321</td>
<td>237,255</td>
<td>237,255</td>
<td>–6%</td>
</tr>
<tr>
<td>3 min</td>
<td>456,822</td>
<td>456,822</td>
<td>456,822</td>
<td>0%</td>
</tr>
<tr>
<td>6 min</td>
<td>237,230</td>
<td>217,851</td>
<td>217,851</td>
<td>–26%</td>
</tr>
</tbody>
</table>

### Table 8
Performance results of the Line Splitting Algorithm variants on Seville underground for experiments with 3 lines under construction.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>(W_{\text{min}})</th>
<th>LSA</th>
<th>Goal</th>
<th>LSA with incremental demand load</th>
<th>LSA with fixed demand load</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10^3)</td>
<td>(z_{\text{GOL}})</td>
<td>(T_{\text{BS}})</td>
<td>(T_{\text{PROV}})</td>
<td>(T_{\text{TOTAL}})</td>
<td>(z_{\text{GOL}})</td>
</tr>
<tr>
<td>3 min</td>
<td>2413</td>
<td>29</td>
<td>211</td>
<td>240</td>
<td>7238</td>
</tr>
<tr>
<td>6 min</td>
<td>2413</td>
<td>25</td>
<td>212</td>
<td>237</td>
<td>7238</td>
</tr>
<tr>
<td>12 min</td>
<td>2578</td>
<td>55</td>
<td>641</td>
<td>696</td>
<td>6024</td>
</tr>
<tr>
<td>3 min</td>
<td>3963</td>
<td>45</td>
<td>152</td>
<td>197</td>
<td>5944</td>
</tr>
<tr>
<td>6 min</td>
<td>7656</td>
<td>49</td>
<td>586</td>
<td>635</td>
<td>8905</td>
</tr>
<tr>
<td>12 min</td>
<td>5477</td>
<td>181</td>
<td>432</td>
<td>613</td>
<td>5477</td>
</tr>
</tbody>
</table>

1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
Table 9
Performance results of the Line Splitting Algorithm variants on Santiago underground for experiments with 3 lines under construction.

<table>
<thead>
<tr>
<th>Wmax</th>
<th>LSA</th>
<th>Goal</th>
<th>LSA with incremental demand load</th>
<th>LSA with fixed demand load</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>It.</td>
<td>pax</td>
<td>op</td>
<td>z_E=0</td>
<td>t_E</td>
</tr>
<tr>
<td>1</td>
<td>152,274</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>304,548</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>456822</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>67,758</td>
<td>1436</td>
<td>2244</td>
<td>3680</td>
</tr>
<tr>
<td>6</td>
<td>131,328</td>
<td>5242</td>
<td>3007</td>
<td>8249</td>
</tr>
<tr>
<td>3</td>
<td>217851</td>
<td>1610</td>
<td>210</td>
<td>1820</td>
</tr>
<tr>
<td>Total</td>
<td>8872</td>
<td>64534</td>
<td>73406</td>
<td>961</td>
</tr>
</tbody>
</table>

Table 10
In-depth analysis of the solution provided by the goal programming model on the real-world networks.

| Network | | Wmax | Stage 1 (min z:pax) | Stage 2 (min z:op) | Reduced operator | Savings |
|---------|-----------------------------|----------------|---------------------|------------------|----------|
| 3       | | 7238 | 7598 | 7238 | 6158 | Veh. setup | 19% |
| 1       | | 7238 | 7598 | 7238 | 6024 | Veh. setup | 21% |
| 12      | | 7238 | 7598 | 7238 | 6024 | Veh. setup | 21% |
| 2       | | 5944 | 14,783 | 5944 | 13,583 | line freq. | Veh. setup | 8% |
| 12      | | 5944 | 14,916 | 5944 | 13,223 | line freq. | Veh. setup | 7% |
| 3       | | 5477 | 17,913 | 5477 | 16,712 | line freq. | Veh. setup | 5% |
| 6       | | 5477 | 17,680 | 5477 | 16,232 | line freq. | Veh. setup | 8% |
| 12      | | 5477 | 17,054 | 5477 | 16,153 | line freq. | Veh. setup | 5% |
| 3       | | 456,822 | 456,822 | 456,822 | 0 | – | 0% |
| 1       | | 264,532 | 820 | 264,532 | 820 | – | 0% |
| 12      | | 264,532 | 820 | 264,532 | 820 | – | 0% |
| 2       | | 237,255 | 1312 | 237,255 | 1304 | line freq. | 2% |
| 12      | | 237,255 | 1312 | 237,255 | 1304 | line freq. | 2% |
| 3       | | 456,822 | 0 | 456,822 | 0 | – | 0% |
| 6       | | 217,851 | 1575 | 217,851 | 1554 | line freq. | 1% |
| 12      | | 217,851 | 1574 | 217,851 | 1574 | – | 0% |

Finally, Table 10 provides an in-depth analysis of the solution provided by the LGP-NDFS model, showing the savings in operator costs when optimized stage 2 on the real-world networks. For each experiment, it is reported the goal function value found for the passenger waiting time (z:pax) and the operator costs (z:op) in the two stages of the LGP-NDFS, the reduced operator costs in stage 2, and the total savings.

The results show that passenger waiting times remain constant when changing the maximum passenger waiting time, except for experiments with Wmax = 3 min in Santiago de Chile network where neither an expansion of the network is done nor passengers use the operating network. However, optimized operator costs decrease moderately as the maximum passenger waiting time increases. Savings in operator costs are only significant in the Seville and Santiago network where operator costs are reduced up to 21% and on average 12%. These savings increase as less lines are constructed and the maximum passenger waiting time increases. The reduced item cost is only related to frequency operation i.e., vehicle and frequency setting costs, thus infrastructure construction and vehicle acquisition costs remain constant. In the Santiago de Chile network, savings are either negligible (2% at most) or nonexistent because the amount of demand to be served does not allow finding alternative least costly planning settings.

8. Conclusions and further research

This work presents an optimization model that integrates the transit network design (TND) and transit frequency setting (TFS) phases in the context of a railway rapid transit system. On the one hand, the TND determines whether the current set of operating lines will be extended. If so, the new lines will be constructed using a pool of candidate corridors and according to the type of service to be performed, without exceeding the available infrastructure budget. On the other hand, the TFS assigns vehicles

Please cite this article as: F. López-Ramos et al., Integrated approach to network design and frequency setting problem in railway rapid transit systems, Computers and Operations Research (2016), http://dx.doi.org/10.1016/j.cor.2016.12.006
and services to the lines while meeting link and vehicle capacity constraints as well as the requirements for vehicle fleet size.

The benefits of this integration are threefold. First, the infrastructure decisions are taken accordingly to the type of services that will be provided, including quality aspects such as the level of service that will actually be provided to passengers. Second, the investments to be carried out consider every possible operation aspect and costs, maximizing resource utilization. Third, operational aspects such as frequency and vehicle acquisition are considered for new and operating lines simultaneously, which once again maximizes the value of the investments.

The optimization model is solved by combining the Lexicographic Goal programming (LGP) technique and a Line Splitting Algorithm (LSA). On the one hand, the LGS optimizes two opposite objectives, minimize passenger riding time and minimize operator costs, giving the priority to passenger goal. Reported results show that significant savings on operational costs can be reached while seeking for an alternative equivalent solution for passenger riding time, if that exists. Thus, the provided solution is more interesting for transport operators. On the other hand, the LSA decomposes the LGP model into smaller problems where only one line is candidate to be constructed. This approach permits solving real-world problems such as the Santiago de Chile and Seville underground networks in reasonable time.

To the best of our knowledge, several existing and new features that had been neither considered nor proposed were successfully applied.

Regarding further research, the authors are working on both computational and modelling aspects, in order to enhance the performance and to broaden the applicability of this transit network design procedure. To speed-up the resolution of the model, a specialized Benders Decomposition [56] might be useful, considering as a subproblem the passenger assignment to the constructed/expanded network provided by the Master problem. Regarding modelling issues, one important contribution is the inclusion of a mode choice model based on transportation mode utilities. That approach can be integrated into an optimization problem using demand mode splitting constraints [57]. Another important aspect is the consideration of passenger waiting time at stations. It is widely thought that its inclusion is rather complex when considering line capacity constraints, as it entails working with a bi-level program. In this bi-level program, the upper level represents the decisions of the planner and the lower level represents decisions of the users [32]. Also, a limitation of the model is that it cannot reflect passenger queueing times at stations due to congestion effects because integrating a congested transit assignment model would increase considerably the model’s complexity, although Codina [58] has recently demonstrated that congested transit assignment models can be reformulated as a variational inequality (VI). These aspects would be required for adapting this model to the design of Bus Rapid Transit Systems. Thus, in its current state, it is only suitable for metro or suburban rail for situations of moderate congestion. The use of the model for other transit systems such as Personal Transit Systems or Freight Transit Systems would require also substantial changes as these systems are intended for serving a point-to-point oriented trip pattern. Finally, some measures of robustness can also be introduced into an optimisation framework as in [41]. Last but not least, different service patterns can be determined according to the period of day in which the demand is given [33].

Acknowledgments

Special thanks to the efforts of two anonymous referees who have contributed to improve the quality and presentation of this paper. Research was supported under Spanish Research Projects TRA2014-52530-C3-3-P and TRA2014-52530-C3-1-P.

References


