Improving robustness of rolling stock circulations in rapid transit networks

Luis Cadarso [a], Ángel Marín [b]

[a] Rey Juan Carlos University, Departamental III Building, Camino del Molino s/n, Fuenlabrada 28943, Spain
[b] Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Aeronáuticos, Plaza Cardenal Cisneros, 3, 28040 Madrid, Spain

ABSTRACT

The rolling stock circulation depends on two different problems: the rolling stock assignment and the train routing problems, which up to now have been solved sequentially. We propose a new approach to obtain better and more robust circulations of the rolling stock train units, solving the rolling stock assignment while accounting for the train routing problem. Here robustness means that difficult shunting operations are selectively penalized and propagated delays together with the need for human resources are minimized. This new integrated approach provides a huge model. Then, we solve the integrated model using Benders decomposition, where the main decision is the rolling stock assignment and the train routing is in the second level. For computational reasons we propose a heuristic based on Benders decomposition. Computational experiments show how the current solution operated by RENFE (the main Spanish train operator) can be improved: more robust and efficient solutions are obtained.

1. Introduction

The globally growing demand and the inability to build more transport infrastructures to increase capacity in most of the world have led to a problem of severe congestion of urban and suburban areas. Congestion threatens our ability to get people where they need to be, with severe economic impacts; it also results in delays that also contribute to negative environmental impacts due to emissions resulting from inefficient system performance: the impacts pose an economic and health threat. Demand for transportation is increasing, so this threat is not going away and must be addressed. Many cities round the world have constructed railway Rapid Transit Networks (RTN) to improve their transport system performance. Our proposal is to meet with the growing demand through improved design and operation of rail RTN. Underground and suburban rail RTN problems are known as high-density network problems, in which the distances between the stations are relatively short and the frequencies are high.

Planning processes related to railway systems are fields that are rich in combinatorial optimization problems. Well-known examples of these are strategic and tactical problems addressed during the planning process. Due to the tremendous size of the planning process, it is usually divided into several steps such as network design, line planning, timetabling and rolling stock (RS) scheduling (i.e., rolling stock assignment and train routing (TR)) [17,18,6]:

1. *Network design*: Designing a RTN is a vital strategic subject due to the fact that it reduces the future traffic congestion, travel time and pollution. The location decisions and the maximum coverage of the demand for the new network are the main goal.
2. *Line planning*: The following step after designing a RTN is planning its lines (origin and destination stations, stops and frequencies). The problem of designing a line system aims at satisfying the travel demand while maximizing the service towards the passengers or minimizing the operating costs of the railway system.
3. *Timetabling*: The general aim of the railway timetabling problem is to construct a train schedule that matches the frequencies determined in the line planning problem.
4. *RS scheduling*: RS circulations are determined once the RS assignment and TR problems are solved:
   - Given a train fleet and finding the optimal composition assignment to each train to satisfy both the timetable and the demand in a dense RTN is known as the rapid transit RS assignment problem; shunting operations are also taken into account in this phase.
   - The train routing (TR) problem is the process of determining a sequence for each train unit in the network once the RS assignment is known. The goal is to obtain sequences that minimize some cost such as the propagated delay in order to achieve a robust solution; a different objective might be

to maximize maintenance opportunities. Train routing planning must allow for each train unit to undergo different types of maintenance checks requirements. However, in our case light maintenance is done during valley hours, and train units that require maintenance are assigned to sequences with maintenance opportunities during valley hours (i.e., train units are swapped at the beginning of the planning period). We assume that the fleet size is large enough to remove any train unit requiring heavy maintenance from the network (this assumption holds true for our case study network).

Traditionally, this planning process has been solved within a hierarchical process, i.e. sequentially. However, this approach may lead the system to operate in an inefficient way (i.e., determining the RS assignment and shunting operations without accounting for train units' sequences and the likely delays might produce sub-optimal plans). An integrated approach may increase the flexibility and the robustness degree of the railway system; therefore, in this paper we propose an integrated mathematical model so as to improve the circulations of the RS in rapid transit networks. The word circulation refers to both the RS assignment and the TR problems.

We present a Robust Circulation of the RS Model (RCRSM), which considers the RS assignment and TR problems in an integrated way. Here robustness means that (1) difficult shunting operations, which may produce negative cascading effects in case of malfunction, are selectively penalized so as to be minimized, (2) propagated delays are minimized, which indirectly minimizes the number of train swapping operations and (3) the need for human resources to perform train units' sequences is minimized.

The rest of this paper is organized as follows. A literature overview and our contributions are presented in Section 2. We describe the rolling stock circulation problem in Section 3. In Section 4, the mathematical formulation is presented in detail. Section 5 contains the solution approach based on Benders decomposition. Section 6 shows computational results based on realistic case studies drawn from RENFE. Conclusions and references follow in the next sections.

2. State of the art

Several researchers have dealt with railway industry planning and managing problems. Alfieri et al. [1] propose an integer programming model so as to determine the RS circulation for multiple RS material types on a single line and on a single day; they use the concept of a transition graph to deal with the RS circulations; this concept is based on the assumption that for each trip, the next trip is known a priori. The objective is to minimize the number of train units such that the given passenger demand is satisfied. The approach is tested on real-life examples from *Nederlandse Spoorwegen* (NS), the main operator of passenger trains in the Netherlands. The model described by Alfieri et al. [1] is extended by Fioole et al. [16], to include combining and splitting trains, as happens at several locations in the Dutch timetable. They use an extended set of variables to locally obtain an improved description of the convex hull of the integer solutions. Robustness is considered by counting the number of composition changes. Maróti [23] focuses on planning problems that arise at NS. He identifies tactical, operational and short-term rolling stock planning problems and develops operations research models for describing them. Peeters and Kroon [26] describe a model and a branch-and-price algorithm to determine a railway rolling stock circulation on a set of train lines. Given the timetable and the passengers' seat demand, the model determines an allocation of rolling stock to the daily trips. They evaluate the solution on three criteria: the service to the passengers, the robustness, and the cost of the circulation. Cadarso and Marín [8] propose a mixed integer optimization model to study suburban rapid transit robust RS assignment. They minimize total costs including service trips, robustness-relevant empty train movements and composition change costs. Almeida et al. [2] state that robustness can be improved by reducing the propagation of delays and increasing the number of feasible resource allocation exchanges. Cadarso and Marín [7] present an integer programming model to determine a sequence of operations to be rolled by the train units such that each operation is included exactly in one sequence and there is always the number of necessary train units available for every operation execution. Cacchiani et al. [5] describe a two-stage optimization model for determining robust rolling stock circulations for passenger trains. They also use the concept of a transition graph. Here robustness means that the rolling stock circulations can better deal with large disruptions of the railway system. They evaluate their approach on the real-life rolling stock-planning problem of NS.

All the previous research address the railway planning problems in a sequential manner. Even the authors who study RS circulations employ the concept of transition graph, which assumes that for each trip, the next trip is known. However, sequential solving approach has many drawbacks [4]. Although practical, the sequential nature of rolling stock assignment and routing optimization leads to sub-optimal plans, with potentially significant economic losses. Improved plans can be generated by building and solving integrated models of some of the planning problems. The airline industry has been a leader in the development of integrated approaches for schedule planning and recovering from disruptions. There has been research in the integration of problems such as flight schedule and fleet assignment [20,9,11], fleet assignment and aircraft routing [25], aircraft routing and crew scheduling [24], and scheduling and competition [27,13]. All these problems were first developed and solved in a sequential fashion. However, the integration of them has outperformed sequential approaches. This fact is demonstrated in every cited paper.

Cadarso and Marín [10] and Cadarso et al. [12] demonstrate that this fact also applies in the railway industry where the integrated timetable planning and RS assignment and integrated disruption management are studied, respectively. Marín et al. [22], López et al. [21] and Walker et al. [29] also develop integrated approaches within the railway industry. Marín et al. [22] and López et al. [21] deal with the integration of railway network design and line planning problems while Walker et al. [29] deal with the simultaneous disruption recovery of a train timetable and crew roster.

Contributions: As we have showed before, the RS assignment and TR problems have been traditionally solved in a sequential fashion. As this sequential solving approach may lead to inefficient (and even infeasible) solutions, we propose a new comprehensive approach to determine optimal train circulations, which integrates RS assignment and TR decisions. The major contributions of this paper include:

1. Development of an integrated schedule optimization model that includes rolling stock assignment and train routing decisions. We avoid using the concept of transition graph, thus developing a more comprehensive decision support tool.
2. Introduction of robustness in the integrated model through different approaches:
   - Penalization of difficult shunting operations, which minimizes negative cascading effects in case of malfunction.
   - Minimization of propagated delays, which means that the number of swapping operations are minimized obtaining robust and improved RS circulations.

- Minimization of the need for human resources to perform train units' sequences.

Altogether, we establish a connection between integration of planning phases and robustness.

3. Definition of a heuristic based on the Benders decomposition so as to improve sequentially obtained circulations in reasonable computational time.
4. Development of case studies using realistic problem instances obtained from the network of the Spanish train operator RENFE; we evaluate several scenarios involving different lines and conduct several sensitivity analysis on various model parameters and a multiobjective optimization analysis.

3. Rolling stock circulation problem description

In this section, the RS circulation problem is described in detail. First, the railway infrastructure is introduced. Then, timetabled services are introduced. Rolling stock and shunting operations are also briefly described. Explanation of the necessary steps to build RS circulations follows. Finally, the passenger demand treatment and the used robustness concepts are described.

3.1. Railway infrastructure

The railway network consists of tracks and stations. Depot stations form a subset of the stations, these are the locations where trains are parked or shunted. We model the infrastructure as a graph with nodes $s \in S$ representing the stations during a determined time period. Depot stations are represented by the subset $SC \subset S$.

Tracks are the existing infrastructure between stations. Between two stations, two different linking infrastructures (i.e., tracks) exist, one for each direction of movement (this assumption holds true for our case study network). We represent the combination of the linking infrastructure and time periods by arcs $a \in A$; every arc is defined by its departure and arrival station and by its departure time.

The railway infrastructure problem (i.e., station and rail corridor location problems) is a strategic problem. Therefore, it is out of the scope of this paper. For comprehensive surveys on this topic see [19] and [28].

3.2. Timetable

The services are grouped in lines. A line is characterized by its terminal stations, by a path through the infrastructure between the terminals, and by a set of stations along the path. Train services run up and down between the terminals and call at the specified stations underway.

The timetable departure times and frequencies are fixed and publicly available. The passengers know when the trains depart and plan their traveling accordingly. Departure times are very inflexible because the time slots are negotiated with a third party (the infrastructure manager) since the network infrastructure is shared among different lines.

We distinguish two types of services: train services represented by $\ell \in L^t$ and empty services represented by $\ell \in L^e$. The set of services is represented by $L = L^t \cup L^e$.

A train service is a passenger train traveling from a depot station to another depot station stopping at a number of intermediate stations. They are characterized by their departure depot station, their arrival depot station, and every arc they travel on, given by $a \in A_\ell \subset A$. The distance rolled by a train service $\ell \in L^t$ is the sum of the lengths of the arcs used by the train service. Train service timetabling is out of the scope of this paper. Therefore, all the train services in the timetable (which has been previously defined) must be performed.

Rapid transit networks are characterized by high frequencies and a lack of capacity in depot stations. These facts make it difficult to operate the network without empty services $\ell \in L^e$. These are defined by an origin, a destination and a departure time. Empty services can help satisfy both capacity and rolling stock material availability in depot stations. For our case study network empty services are not given in the timetable. Therefore, we will consider a finite set of empty services and the model will choose between this pre-fixed set of options.

3.3. Rolling stock and shunting

There are self-propelled train units of type $m \in M$; they all have a driver seat at both ends. Units of the same type can be attached to each other to form trains compositions. A composition $c \in C$ of train units is a sequence of elements of M. For our case study network, train compositions are predefined: compositions of up to two train units are allowed.

Shunting operations complicate rapid transit networks because the performance time is on the order of the service frequency time. They are only performed in depot stations. Train units of the same type can be aggregated to form longer compositions, and compositions can be disaggregated into individual train units. Although composition changes enable the network operator to use smaller fleet sizes, it is always a complicating operation, due to the necessity of human resources and the possibility of failure in the mechanical system governing the process.

3.4. Determining rolling stock circulations

The RS circulations are completely determined with the RS assignment and the TR problems.

Rolling stock assignment: The goal of the RS assignment problem is to determine the type and number of train units to be assigned to the services considering a given timetable and a demand to satisfy in a context in which shunting is optimized.

Train routing: The TR problem aims at assigning each individual train unit, referred to as an identification number, to RS operations (i.e., services, aggregations and disaggregations). Given the RS assignment, we must determine a sequence of operations to be rolled by an individual train unit such that each operation is included in exactly one sequence.

3.5. Passenger demand

For passenger demand, we use the expected number of passengers using each service, which is given by RENFE. The passenger demand for this problem is treated as a passenger flow $pf_{a,\ell}$ through each space-time arc a belonging to each train service $\ell \in L^t$. This passenger flow is obtained under normal conditions (i.e., assuming that the train services matched the designed timetable). Under this hypothesis, the model will treat the passengers from a centralized point of view (i.e., only the operator criteria are optimized). However, since the proposed problem relates to a suburban rapid transit network, it is obvious that every passenger will have the option to choose any other available company or transportation mode. Thus, the operator has to factor in passenger behavior to avoid losing passengers to other transportation companies. In a simplified approach, passenger behavior can be summarized as follows: if the passenger maintains his/her satisfaction with the transportation mode, he/she will remain in the system. As long as the system operator maintains certain standards within the transportation system, we assume that the passenger flow is known.

Under the assumption that public timetables are met, transportation standards might be described by the capacity offered in each train service. The composition assigned to each train service will be a tradeoff between the operating costs and the behavior of passengers (represented by their comfort level).

We acknowledge that this approach treats the demand heuristically. The approach is unable to trace individual passengers; instead, it considers demand on the arcs (i.e., between successive stations). Passengers on a longer journey appear in the demand of each arc underway. Whenever the demand of an arc exceeds the allocated capacity, part of the demand remains unattended: these passengers are unattended, and they are supposed to leave the system. However, the demand on successive arcs is not linked to each other. Therefore, an unattended passenger still shows up in the demand of later arcs. This approach is heuristic in that it ignores the dynamic interaction between demand and supplied capacity. However, Cadarso et al. [12] justify this heuristic approach.

Elasticity of demand is out of the scope of this paper. However, level of service, which might be perceived through attributes such as price of travel and total trip time, affects the total volume of demand [14]. Nevertheless, we assume that the demand is fixed. This assumption is reasonable when the entry of new operators is unlikely. However, if new operator entry occurs, the unconstrained demand will likely change due to elasticity of demand. Cadarso et al. [13] account for elasticity of demand as a result of new services.

3.6. Robustness

Robustness is introduced penalizing composition changes and empty services. When a composition change is performed, multiple failures can occur, forcing the train unit to be parked for a long time and causing a disruption. The mechanical system used to perform a composition change sometimes fails and requires extra time to enact the change. Above all, composition change times are overestimated to account for the effects mentioned above and to introduce robustness into the system. Finally, if a malfunction occurs it must be contained to avoid negative disruptive effects. Containment is easier if the disruptive event occurs during off-peak hours. Similarly, empty services during rush hours complicate network operation because they use the same infrastructure as train services. During rush hours they are also heavily penalized. It is also better to avoid (if possible) empty services to destination depot stations with time-dependent capacities (i.e., stations that are shared with different lines). The system is made more robust by assigning only one material type per line (i.e., for every train service operating the same line, the material must be equal). This constraint increases swapping opportunities between different train services at depot stations in the same line. Thus, propagation of negative disruptive effects may be mitigated in an easier way. This assumption holds true for our case study network. However, it could be easily relaxed in our modeling approach.

Delay penalization is another way to include robustness to the model. The propagated delay from one operation (i.e., services, aggregations and disaggregations) to another operation (operation connection) is defined by $pd = \max(ad - slack, 0)$, where ad is an aleatory variable representing the arrival delay (see [7]), and $slack$ is the planned slack between both operations. When $pd > 0$, the train unit performing the first operation will not be on time to perform the following operation, that is, the operation connection cannot be performed. However, the operation connection will be considered feasible but penalized in the objective function. This feasibility is justified for RENFE performance, because in real operation, they do not permit propagated delays. Although there is a lack of capacity and resources in the network, RENFE planners always reserve some train units and, in case of delays, they swap to avoid delay propagation. Consequently, the obtained statistical data might be considered independent of the sequence. Thus, this robustness criterion indirectly minimizes the number of necessary swapping operations and the human resources required to perform the material swapping.

Another issue affecting train units' performance is their ability to be ready for departure. Train units are equipped with air compressed circuits to activate brakes. When a train unit is stopped, these circuits are automatically emptied and some time is needed to inflate them again. Therefore, in order to prepare a train unit (which has been stopped for a while) for departure, human resources are needed. Thus, robustness may be also introduced through penalizing crew requirement due to brake malfunction between two services in a sequence. With this criterion more human resources will be available in case of disruptive events and their negative effects will be contained in an easier way.

4. Robust circulation of the rolling stock model

Railway planning is currently divided into several optimization steps from first strategic decisions to daily operations. It is well known that disintegrated planning produces optimal solutions for each stage but non-optimal global solutions. We propose a new integrated model for the RS and TR problems so as to obtain better train circulations. A greater robustness degree and more efficient schedules may be obtained through an integrated approach, which considers two consecutive planning stages: one way of obtaining a high-quality solution in the second stage is the introduction of some *slack* in the first stage. Similarly, a robust solution adds *slack* to safeguard against data perturbation. This slack may provide a *smooth interface* between subsequent planning stages. This fact becomes a connection between robustness and integration.

The Robust Circulation of the RS Model (RCRSM) is based on the rolling stock assignment and train routing models proposed by Cadarso and Marín [7,8]. They considered the railway rolling stock and train routing problems for rapid transit networks. Compared to these two papers, the novelty of the current paper lies in the following aspects:

- The development of a multiobjective integrated linear integer programming model for schedule optimization that combines RS assignment and TR problems.
- The development of an algorithmic framework that allows us to find solutions to the integrated model, which are demonstrated to be superior to the ones obtained with the sequential approach presented in [8,7].
- We solve this model using realistic problem instances obtained from the network of the Spanish railway operator RENFE. We also perform sensitivity analysis on model parameters and conduct a multiobjective optimization study.

In order to be able to formulate the RCRSM, we need to define the following sets, parameters and variables.

4.1. Sets

- L is the set of services indexed by ℓ. Each service is characterized by an origin, a destination and a departure time.
- $L^t \subset L$ is the subset of train services. These services are performed to attend the passenger demand.
- $L^e \subset L$ is the subset of empty services. These services cannot attend passenger demand.

- S is the set of nodes indexed by s and s'. The nodes are defined by a station and a time period. s^- denotes the preceding node to s (i.e., s^- and s represent the same station during consecutive time periods).
- $SC \subset S$ is the subset of depot nodes.
- $CS \subset S$ is the subset of nodes at which the material is counted.
- A is the set of arcs indexed by a. They are characterized by a departure station and time period and an arrival station.
- M is the set of train unit material types indexed by m.
- C is the set of compositions indexed by c.
- $C_m \subset C$ is the subset of compositions composed of material type m.
- A_ℓ is the set of arcs a served by train service $\ell \in L^t$.
- I is the set of operations indexed by i and j. There are three types of operations: train and empty services (indexed by 1), aggregations (indexed by 2) and disaggregations (indexed by 3); the last two types of operations refer to composition changes.

4.2. Parameters

- oc_c is the operating cost per rolled kilometer for composition c.
- km_ℓ is the number of kilometers in service ℓ.
- $upc_{a,\ell}$ is the unattended passenger cost in arc a served by train service $\ell \in L^t$.
- ϑ_s is the penalty for composition change in node s.
- ψ, ζ are the train delay cost per time period and the need for human resources cost per time period, respectively.
- $E[pd_{i,s}^{j,s'}]$ is the number of expected delay time periods propagated from operation i ending at node s to operation j beginning at node s'.
- $cr_{i,s}^{j,s'}$ is the penalty for the need for human resources to enable train units to be ready between operation i ending at node s and operation j beginning at node s'.
- $pf_{a,\ell}$ is the expected passenger flow in arc a served by train service $\ell \in L^t$.
- q_c is the passenger capacity in composition c.
- cap_s is the depot station s capacity.
- $\alpha_{\ell,s} = -1(1)$, if service ℓ leaves from node s (if service ℓ arrives at node s).
- χ_m is the fleet size for train unit material of type m.
- e, d are the time needed for aggregation and disaggregation, respectively.
- tu_c is the number of train units in composition c.
- $\beta_\ell = 1(0)$, if service ℓ is operating at the count time period (otherwise).
- $cn_{c,c'}$ is the number of compositions c needed to obtain a composition c' in case of aggregation (number of compositions c' obtained from composition c in case of disaggregation).
- $\mu_{s',s} = 1$, if a composition change which started in node s' is still being operated in node s.
- s_i, s_f are the initial and final nodes in the planning period.
- $\tilde{\alpha}_{i,c} = 2(1)$, if i is a disaggregation (i.e., $i=3$) followed by trains with c composition (otherwise).
- $\tilde{\beta}_{j,c} = 2(1)$, if j is an aggregation operation (i.e., $j=2$) preceded by trains with c composition (otherwise).
- ω is a scalar ranging in [0, 1]. Its different values produce different Pareto optimal solutions.

4.3. Variables

The most central decision variables are $x_\ell^c \in \{0,1\}$ defined for $\ell \in L, c \in C$ and $seq_{i,s}^{j,s',c} \in \{0,1\}$ defined for $i,j \in I, s, s' \in SC, c \in C$:

- x_ℓ^c variables take value 1 so as to indicate whether composition $c \in C$ is scheduled for service $\ell \in L$.
- $seq_{i,s}^{j,s',c}$ variables take value 1 if operation i ending at node s is followed by operation j beginning at node s', both of them with composition c.

The model contains the following additional variables:

- y_s^c are non-negative integer variables. They denote the train inventory of composition c in node s.
- $up_{a,\ell}$ are non-negative variables. They denote unattended passengers in arc a served by train service $\ell \in L^t$.
- $\epsilon_s^{c,c'}(\delta\epsilon_s^{c,c'}) \in \{0,1\}$. They take value 1 if an aggregation (a disaggregation) starts at node s, from composition c to composition c'.
- $cc_s^{c,c'} \in \{0,1\}$. They take value 1 if a composition change starts at node s, from composition c to composition c'.
- $\phi_{i,s}^c(\varphi_{j,s}^c)$ are non-negative variables. They determine the number of operation i that end (j that begin) in node s with composition c; 0, otherwise.

Note that $\phi_{i,s}^c$ and $\varphi_{j,s}^c$ are the only variables that link the rolling stock assignment to the train routing constraints. Therefore, the proposed integrated model can be used for any underlying rolling stock scheduling or train routing problems as long as they are expressed in terms of $\phi_{i,s}^c$ and $\varphi_{j,s}^c$.

The RCRSM mathematical formulation is as follows.

4.4. Objective function

$$\min z = \omega\left[\sum_{\ell \in L}\sum_{c \in C} oc_c km_\ell x_\ell^c + \sum_{\ell \in L^t}\sum_{a \in A_\ell} upc_{a,\ell} up_{a,\ell} + \sum_{s \in SC}\sum_{c,c' \in C} \vartheta_s \cdot cc_s^{c,c'}\right] + (1-\omega)\left[\sum_{i,j \in I}\sum_{s,s' \in Sc}\sum_{c \in C}\left[\psi \cdot E\left[pd_{i,s}^{j,s'}\right] + \zeta \cdot cr_{i,s}^{j,s'}\right] \cdot seq_{i,s}^{j,s',c}\right]. \tag{1}$$

We minimize the following items in the objective function:

1. Operating costs for train and empty services (i.e., $\ell \in L$); for *dangerous* empty services $\ell \in L^e$ the value km_ℓ is increased in order to introduce robustness.
2. Unattended passengers costs, that is, every passenger who cannot board the train due to capacity shortages is assumed to be a cost for the operator.
3. Shunting costs; the parameter ϑ_s is modulated depending on the shunting operation, that is, if a shunting operation is likely to fail, the associated cost to it will be increased to introduce robustness.
4. Expected delay; for calculating the $pd_{i,s}^{j,s'}$ values, expressed in minutes, we take advantage of the fact that the slack between operations is always a constant value. Therefore, this value can be captured through the location parameter $\theta_{i,s}^{j,s'}$, and the propagated delay is calculated in (2) (see [7]):

$$E[pd_{i,s}^{j,s'}] = \theta_{i,s}^{j,s'} \cdot \left[1 - \phi\left(\frac{\ln\left(\frac{\gamma - \theta_{i,s}^{j,s'}}{m_{i,s}}\right)}{\sigma_{i,s}}\right)\right] + m_{i,s} \cdot e^{\sigma_{i,s}^2/2} \cdot \left[1 - \phi\left(\frac{\ln\left(\frac{\gamma - \theta_{i,s}^{j,s'}}{m_{i,s}}\right)}{\sigma_{i,s}} - \sigma_{i,s}\right)\right], \tag{2}$$

where $\phi(x)$ is the cumulative distribution function of a standard normal distribution, $m_{i,s}$ is the scale factor, $\sigma_{i,s}$ is the standard deviation, and γ is equal to 0 if $\theta_{i,s}^{j,s'} \le 0$ or γ is equal to $\theta_{i,s}^{j,s'}$, otherwise.

5. The need for human resources to enable train units to be ready between operations; the penalty $cr_{i,s}^{j,s'}$ depends on the time the train unit has been stopped. First, it grows linearly with time, and after a while it is constant. Consequently, a piecewise penalization function is used for $cr_{i,s}^{j,s'}$.

Even though the presented objective function is a cost function (its units are monetary units), we acknowledge that there are two main parts of different nature: rolling stock assignment costs, which are given by the first three terms, and train routing costs, which are given by the last two terms (these ones are transformed in monetary units with ψ, ζ). Therefore, the presented problem is a multiobjective problem, which is formulated as a single-objective optimization problem. The optimal solutions to the single-objective optimization problem are Pareto optimal solutions to the multiobjective optimization problem (i.e., different values of ω produce different Pareto optimal solutions).

The robust circulation of the rolling stock model (RCRSM) described here is formulated as a multicommodity flow model. It minimizes a combination of system-related and service-related criteria subject to constraints for the underlying rolling stock assignment and train routing problems. The purpose of the constraints is summarized as follows:

- The passenger demand for each train service is linked to the capacity of the allocated train units.
- As for the rolling stock, each service gets at most one composition assigned; train units' flow conservation is ensured; the amount of used rolling stock is limited; the storage and shunting capacity of the stations is controlled.
- Coupling constraints establish the relationship between rolling stock constraints and train routing constraints.
- For train routing, each operation gets a predecessor and a successor operation (sequence constraints).

4.5. Passengers constraints

The following group of constraints links the allocated seat capacity to the number of passengers $pf_{a,\ell}$:

$$\sum_{c \in C} q_c x_\ell^c \geq pf_{a,\ell} - up_{a,\ell} \quad \forall \ell \in L^t, \ a \in A_\ell \tag{3}$$

Constraints (3) say that for each arc $a \in A$ attended by train service $\ell \in L^t$, the capacity of the train is enough to accommodate the passenger demand minus the number of unattended passengers.

4.6. Rolling stock constraints

$$\sum_{c \in C} x_\ell^c = 1 \quad \forall \ell \in L^t \tag{4}$$

$$\sum_{c \in C} x_\ell^c \leq 1 \quad \forall \ell \in L^e \tag{5}$$

Constraints (4) state that each train service $\ell \in L^t$ must be assigned a composition c. Constraints (5) express that empty services $\ell \in L^e$ get at most one composition:

$$\begin{aligned} y_{s^-}^c &+ \sum_{\substack{\ell \in L \\ \alpha_{\ell,s}=1}} x_\ell^c + \sum_{c' \in C} \epsilon_{s^{-e}}^{c',c} + \sum_{c' \in C} cn_{c',c} \cdot \delta\epsilon_{s^{-d}}^{c',c} \\ &= y_s^c + \sum_{\substack{\ell \in L \\ \alpha_{\ell,s}=-1}} x_\ell^c + \sum_{c' \in C} cn_{c,c'} \cdot \epsilon_s^{c,c'} + \sum_{c' \in C} \delta\epsilon_s^{c,c'} \quad \forall s \in SC, \ c \in C \end{aligned} \tag{6}$$

Inventory conservation constraints (6) ensure the train units' flow balance. These constraints consider the increase or decrease of the inventory depending on departing and arriving services and also on local shunting operations ($s^{-e}(s^{-d})$ represents the same station as s but $e(d)$ time periods before). Obviously, the inventory is always non-negative.

$$\begin{aligned} &\sum_{s \in CS} \sum_{c \in C_m} tu_c y_s^c + \sum_{\ell \in L} \sum_{c \in C_m} tu_c \beta_\ell x_\ell^c \\ &+ \sum_{s \in CS} \sum_{s' \in SC} \sum_{c,c' \in C_m} \mu_{s',s}(tu_{c'} \epsilon_{s'}^{c,c'} + tu_c \delta\epsilon_{s'}^{c,c'}) \leq \chi_m \quad \forall m \in M \end{aligned} \tag{7}$$

$$\sum_{c \in C} tu_c y_s^c + \sum_{s' \in SC} \sum_{c \in C} \mu_{s',s}(tu_{c'} \epsilon_{s'}^{c,c'} + tu_c \delta\epsilon_{s'}^{c,c'}) \leq cap_s \quad \forall s \in SC \tag{8}$$

Fleet capacity constraints (7) ensure that the number of train units used at the count time period is limited by the size of the fleet. Note that these constraints count the running trains and those ones in depot stations. Depot capacity constraints (8) ensure that the total capacity of the station is not exceeded:

$$cc_s^{c,c'} = \epsilon_s^{c,c'} + \delta\epsilon_s^{c,c'} \quad \forall s \in SC, \ c, c' \in C \tag{9}$$

$$y_{s_i}^c = y_{s_f}^c \quad \forall s \in SC, \ c \in C \tag{10}$$

Constraints (9) count the number of composition changes in every depot station. Note that for all the composition changes that are not physically possible (i.e., due to composition incompatibility), the variables $\epsilon_s^{c,c'}, \delta\epsilon_s^{c,c'}$ are fixed to zero value. Constraints (10) denote that the inventory during the initial and final nodes must be equal. Consequently, the obtained schedule is periodic.

4.7. Coupling constraints

$$x_\ell^c = \varphi_{1,s}^c \quad \forall \ell \in L, \ c \in C, \ s \in SC : \alpha_{\ell,s} = -1 \tag{11}$$

$$x_\ell^c = \phi_{1,s}^c \quad \forall \ell \in L, \ c \in C, \ s \in SC : \alpha_{\ell,s} = 1 \tag{12}$$

$$\epsilon_s^{c,c'} = \varphi_{2,s}^c \quad \forall s \in SC, \ c, c' \in C \tag{13}$$

$$\epsilon_s^{c,c'} = \phi_{2,s^{+e}}^{c'} \quad \forall s \in SC, \ c, c' \in C \tag{14}$$

$$\delta\epsilon_s^{c,c'} = \varphi_{3,s}^c \quad \forall s \in SC, \ c, c' \in C \tag{15}$$

$$\delta\epsilon_s^{c,c'} = \phi_{3,s^{+d}}^{c'} \quad \forall s \in SC, \ c, c' \in C \tag{16}$$

Constraints (11)–(16) determine the characteristics of each operation: the node at which it starts and ends and the composition assigned. This information is stored in the variables $\phi_{i,s}^c, \varphi_{j,s}^c$. For example, in constraints (11) each time that variable x_ℓ^c takes value 1, the right hand side of the constraint will also be one. Therefore, variables $\varphi_{1,s}^c$ will take value 1 every time a train or empty service departs from s with composition c. Similarly, variables $\phi_{1,s}^c$ will take value 1 every time a train or empty service arrives at s with composition c. The same explanation follows for constraints (13)–(16). $s^{+e}(s^{+d})$ represents the same station as s but e (d) time periods later.

4.8. Sequence constraints

$$\sum_{j \in I} \sum_{s' \in S} seq_{i,s}^{j,s',c} = \tilde{\alpha}_{i,c} \phi_{i,s}^c \quad (\kappa_{i,s}^c) \quad \forall i \in I, \ s \in S, \ c \in C \tag{17}$$

$$\sum_{i \in I} \sum_{s' \in S} seq_{i,s}^{j,s',c} = \tilde{\beta}_{j,c} \varphi_{j,s}^c \quad (\lambda_{j,s}^c) \quad \forall j \in I, \ s \in S, \ c \in C \tag{18}$$

Constraints (17)–(18) are sequencing constraints. They ensure that every operation is preceded by another one and that every operation is followed by other operation. Here, the variables $\phi_{i,s}^c, \varphi_{j,s}^c$ are used. The former introduces i's operation ending information in order to find a following compatible operation. Similarly, the latter shows beginning information of operation j to find a preceding compatible operation. $\kappa_{i,s}^c, \lambda_{j,s}^c$ are dual variables.

The RCRSM presented formulation leads us to a huge model size. For a real instance in the rapid transit network in Madrid, we have with this new integrated formulation more than one hundred million binary variables. In order to deal with such a huge model we employ Benders decomposition.

5. Solution approach: Benders decomposition

Benders decomposition may be obtained by classifying variables into difficult and easy variables. Here, the difficult ones are the RS and coupling variables: $[x^c_\ell, y^c_s, up_{a,\ell}, \epsilon^{c,c'}_s, \delta\epsilon^{c,c'}_s, cc^{c,c'}_s, \phi^c_{i,s}, \varphi^c_{j,s}]^T$. Once difficult variables are known, it is relatively easy to determine the optimal sequence associated to them; easy variables are sequence variables ($seq^{j,s',c}_{i,s}$). Hence, we have that the Benders Master Model (*MM*) will be composed of the passengers constraints (3), RS constraints (4)–(10), coupling constraints (11)–(16) and both a new objective function and Benders optimality cuts to be defined; and the Benders SubModel (*SM*) will be the train routing (TR) model.

We define in the following subsections the Benders *SM* and *MM*.

5.1. Benders submodel

The Benders *SM* at each iteration *it* (SM^{it}) will be the train routing model:

$$\min z = (1-\omega)\left[\sum_{i,j\in I}\sum_{s,s'\in S}\sum_{c\in C}\left[\psi\cdot E\left[pd^{j,s'}_{i,s}\right]+\zeta\cdot cr^{j,s'}_{i,s}\right]\cdot seq^{j,s',c}_{i,s}\right] \tag{19}$$

$$\sum_{j\in I}\sum_{s'\in S} seq^{j,s',c}_{i,s} = \tilde{\alpha}_{i,c}\phi^{c,it}_{i,s} \quad \left(\kappa^c_{i,s}\right) \quad \forall i\in I,\ s\in S,\ c\in C \tag{20}$$

$$\sum_{i\in I}\sum_{s'\in S} seq^{j,s',c}_{i,s} = \tilde{\beta}_{j,c}\varphi^{c,it}_{j,s} \quad \left(\lambda^c_{j,s}\right) \quad \forall j\in I,\ s\in S,\ c\in C \tag{21}$$

where $\phi^{c,it}_{i,s}$ and $\varphi^{c,it}_{j,s}$ come from the *MM* solution at iteration *it* (MM^{it}). Consequently, the right hand side of the train routing constraints becomes a datum.

Therefore, the SM^{it} can be reformulated as in [7], where the authors develop an integer model to determine train sequences once the RS assignment is known. In this case, for each iteration we will know the RS assignment from the *MM*, so we can reformulate the *SM* into the mentioned formulation.

In the reformulated Benders *SM* defined by (22)–(25) indexes i', j' are appearing. They do not have the same meaning as before. In the previous formulation i,j only referred to the type of operation (i.e., services, aggregations or disaggregations). However, i', j' are now numbering operations. For example, a service may be numbered as operation number 56: departure and arrival stations and times as well as assigned composition are associated to this operation number. Consequently, we know whether two different operations are compatible or not by means of the their numbers. This compatibility is given by the new set $CO^{j'}_{i'}$, which elements represent whether operations i', j' are compatible or not.

Therefore, when the index i' appears in the reformulated *SM*, we are actually referring to *i,s,c*. The same applies for j'. Algorithm 1 performs a mapping between the different formulations. Here, $t_{i,s}$ is the time duration of operation *i* which started in node *s*.

Algorithm 1. Mapping from $\tilde{\beta}_{i,c}\varphi^c_{i,s}, \alpha_{i,c}\phi^c_{i,s}$ to $\beta_{i'}, \alpha_{i'}$.

$i' = 1$
for $i = 1$ **to** $|I|$ **do**
  for $c = 1$ **to** $|C|$ **do**
    for $s = 1$ **to** $|S|$ **do**
      if $\varphi^c_{i,s} = 1$ **then**
        $\beta_{i'} \leftarrow \tilde{\beta}_{i,c}\varphi^c_{i,s}$
        $\alpha_{i'} \leftarrow \tilde{\alpha}_{i,c}\phi^c_{i,s+t_{i,s}}$
        $i' \leftarrow i' + 1$
      end if
    end for
  end for
end for
return $\alpha_{i'}, \beta_{i'}$

$$\min z = (1-\omega)\left[\sum_{i'\in I'}\sum_{j'\in I'}\left(\psi E\left[pd^{j'}_{i'}\right]+\zeta cr^{j'}_{i'}\right)seq^{j'}_{i'}\right] \tag{22}$$

Subject to :

$$\sum_{j'\in CO^{j'}_{i'}} seq^{j'}_{i'} = \alpha^{it}_{i'} \quad (\kappa_{i'}) \quad \forall i' \in I' \tag{23}$$

$$\sum_{i'\in CO^{j'}_{i'}} seq^{j'}_{i'} = \beta^{it}_{j'} \quad (\lambda_{j'}) \quad \forall j' \in I' \tag{24}$$

$$seq^{j'}_{i'} \in \mathcal{R}^+ \quad \forall i', j' \in I' \tag{25}$$

We have relaxed the integrality property of the binary variable $seq^{j'}_{i'}$ considering that the relaxed *SM* with integer data has an integer solution (see Proposition 5.1). The TR model formulation used in this work is developed in order to provide feasible sequences whatever the RS assignment is. This is achieved by ensuring minimum idle times between operations. Hence, the routing *SM* will always be feasible and no feasibility cuts are needed for the *MM*.

The Dual model of the Reformulated *SM* at each iteration *it* ($DRSM^{it}$) is as follows:

$$\max z_{DRSM^{it}} = \sum_{i'\in I}\alpha^{it}_{i'}\kappa_{i'} + \sum_{j'\in I}\beta^{it}_{j'}\lambda_{j'} \tag{26}$$

$$\kappa_{i'} + \lambda_{j'} \le (1-\omega)[\psi E[pd^{j'}_{i'}] + \zeta cr^{j'}_{i'}] \quad (seq^{j'}_{i'}) \quad \forall i',\ j' \in CO^{j'}_{i'} \tag{27}$$

$$\kappa_{i'} \in \mathcal{R} \quad \forall i' \in I \tag{28}$$

$$\lambda_{j'} \in \mathcal{R} \quad \forall j' \in I. \tag{29}$$

Proposition 5.1. *Under the assumption of integrality of vector b, the linear programming relaxation* $\min\{z(\boldsymbol{x}) : A\cdot\boldsymbol{x} = b, \boldsymbol{x}\in\mathcal{R}^+\}$ *of the train routing problem* $\min\{z(\boldsymbol{x}) : A\cdot\boldsymbol{x} = b, \boldsymbol{x}\in\mathcal{Z}^+\}$ *will have an optimal solution that is integer.*

Proof. Suppose we have an optimal basis *B* from the reformulated *SM*. From linear programming we know that *B* is a non-singular submatrix of *A*, where *A* is the coefficient matrix. *B* will hold the following attributes:

- it will be always composed of elements belonging to the following set: {0,1} (see constraints (23) and (24)), thus the first condition for *B* being totally unimodular [30] is matched;
- and, due to the problem characteristics (i.e., every operation *j* must be preceded (followed) by a unique one, except for aggregations (disaggregations) where at most two different operations must precede (follow)), *B* will be always totally unimodular: every column in *B* will have at most two non-zero elements, and its rows can be partitioned into two sets such

that two nonzero entries in a column are in different sets of rows.

Hence, the optimal basis B will always be totally unimodular and $\det(B) = \pm 1$ (B is an optimal basis: $\det(B) \neq 0$), so the linear relaxation solves the integer problem [30]. □

In order to build the Benders *MM* optimality cuts, once we know the dual variables $\kappa_{i'}, \lambda_{j'}$ we must obtain the variables in the Benders *MM*: $\kappa^c_{i,s}, \lambda^c_{j,s}$. Algorithm 2 performs a mapping between the variables. Here, $OP_{i'}$ determines the type of operation (i.e., services, aggregations and disaggregations) of i'; $AS_{i'}$ determines the arrival node s of operation i'; $DS_{i'}$ determines the departure node s of operation i'; and $C_{i'}$ denotes the assigned composition c to operation i'.

Algorithm 2. Mapping from $\kappa_{i'}, \lambda_{i'}$ to $\kappa^c_{i,s}, \lambda^c_{i,s}$.

1: **for** $i' = 1$ **to** $|I'|$ **do**
2:   **for** $i=1$ **to** $|I|$ **do**
3:     **for** $c=1$ **to** $|C|$ **do**
4:       **for** $s=1$ **to** $|S|$ **do**
5:         **if** $i \in OP_{i'}$ **and** $c \in C_{i'}$ **and** $s \in AS_{i'}$ **then**
6:           $\kappa^c_{i,s} \leftarrow \kappa_{i'}$
7:         **end if**
8:         **if** $i \in OP_{i'}$ **and** $c \in C_{i'}$ **and** $s \in DS_{i'}$ **then**
9:           $\lambda^c_{i,s} \leftarrow \lambda_{i'}$
10:         **end if**
11:       **end for**
12:     **end for**
13:   **end for**
14: **end for**
15: **return** $\kappa^c_{i,s}, \lambda^c_{i,s}$

5.2. Benders Master Model

The Benders MM^{it} will be as follows:

$$\min z_{MM^{it}} = \omega \left[\sum_{\ell \in L} \sum_{c \in C} oc_c km_\ell x^c_\ell + \sum_{\ell \in L^t} \sum_{a \in A_\ell} upc_{a,\ell} up_{a,\ell} + \sum_{s \in S} \sum_{c,c' \in C} \vartheta_s \cdot cc^{c,c'}_s \right] + \upsilon \tag{30}$$

subject to constraints (3)–(16) and

$$\upsilon \geq \sum_{i \in I} \sum_{s \in S} \sum_{c \in C} \tilde{\alpha}_{i,c} \phi^c_{i,s} \kappa^{c,it}_{i,s} + \sum_{j \in I} \sum_{s \in S} \sum_{c \in C} \tilde{\beta}_{j,c} \varphi^c_{j,s} \lambda^{c,it}_{j,s} \quad \forall it \in AOBC_{it}, \tag{31}$$

where (31) are Active Optimality Benders Cuts at iteration *it* ($AOBC_{it}$).

Algorithm 3 shows the overall Benders decomposition procedure. UB^{it} and LB^{it} are the upper and lower bounds of the solution process, respectively. ϵ is the allowable error. $z^*_{MM^{it}}$ is the objective function value of the master model at iteration *it*. $z^*_{DRSM^{it}}$ is the objective function value of the dual reformulated submodel at iteration *it*. $\upsilon^*, x^{*c}_\ell, seq^{*j'}_{i'}$ are the optimal values for the model variables $\upsilon, x^c_\ell, seq^{j'}_{i'}$.

Algorithm 3. Benders decomposition.

1: **Set** $AOBC_{it} \leftarrow \emptyset$; $UB^{it} \leftarrow \infty$; $LB^{it} \leftarrow -\infty$; $it=1$, ϵ
2: **Solve** MM^{it}
3: $LB^{it} = \max\{LB^{it-1}, z^*_{MM^{it}}\}$
4: **Call** Algorithm 1
5: **Solve** $DRSM^{it}$
6: **Call** Algorithm 2
7: $UB^{it} = \min\{UB^{it-1}, z^*_{MM^{it}} - \upsilon^* + z^*_{DRSM^{it}}\}$
8: $GAP = \frac{UB^{it} - LB^{it}}{LB^{it}}$
9: **if** $GAP < \epsilon$ **then**
10:   **return** $x^{*c}_\ell, seq^{*j'}_{i'}$
11:   **Stop**
12: **end if**
13: $AOBC_{it} \leftarrow yes$
14: $it \leftarrow it+1$
15: **Go to** 2

6. Computational experiments

Our experiments are based on realistic cases drawn from RENFE's regional network in Madrid (Fig. 1), also known as "Cercanías Madrid". Nowadays, RENFE makes its planning in a sequential manner: the circulations are obtained by solving two problems in an isolated way. The two problems are the rolling stock assignment and train routing problems. Once they are solved they obtain train circulations. Furthermore, RENFE does not use operations research techniques to obtain the circulations. Consequently, their schedule may be improved by using them. See [7] and [8] for further details in the application of operations research techniques to RENFE planning: the authors propose a sequential planning. However, sequential planning may lead to a suboptimal schedule. We show here how integrated planning may overcome the sequential approach by producing better, smoother and more robust plans.

The "Cercanías Madrid" rapid transit network is characterized by its modular structure. That is, in real life it is separated into different and independent modules for operating purposes. Every module has its own infrastructure. Here, we study two different modules: the first one composed of line C5, and the second one composed of lines C3 and C4.

Line C5 has 23 stations (Fig. 2) and 4 depot stations: Mostoles el Soto, Atocha, Fuenlabrada and Humanes. It has more than 320 train services scheduled each day with frequencies on the order of 3 min during peak hours. There is one material type available and train services can be assigned compositions of one train unit or two train units.

In lines C3 and C4, there are nearly 400 scheduled train services each day with frequencies on the order of 10 min. Line C3 is composed of 12 stations (Fig. 3) and 3 depot stations: Chamartín, Atocha and Aranjuez. In line C4, there are 18 stations (Fig. 4) and 8 depot stations: Parla, Parla Industrial, Getafe Centro, Atocha, Chamartín, Tres Cantos, Alcobendas and Colmenar Viejo. Some depot stations are shared between both lines. There is one material type available and the train services can be assigned compositions of one train unit or two train units. The same material is used for both lines, so the RS can be interchanged between them.

Our runs are performed on a Personal Computer with an Intel Core2 Quad Q9950 CPU at 2.83 GHz and 8 GB of RAM, running under Windows 7 64Bit, and our programs are implemented in GAMS 23.2.1/Cplex 12.1.

The rest of this section is divided into different subsections: in Section 6.1 we aim at applying the Benders decomposition presented in Section 5; however, this approach fails to provide optimal and even feasible solutions for the operator, thus we propose a heuristic in Section 6.2 so as to polish sequentially obtained solutions. For the study cases in Sections 6.1 and 6.2 we assume that $\omega=0.5$. We relax this assumption in Section 6.3, where we perform sensitivity analysis on various model parameters and conduct a multiobjective optimization study.

Fig. 1. RENFE's rapid transit network around Madrid.

Fig. 2. Line C5.

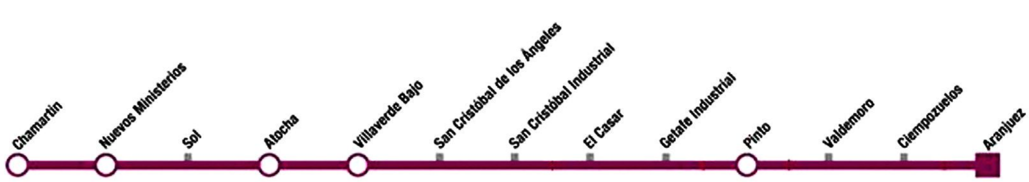

Fig. 3. Line C3.

6.1. Benders decomposition

This subsection shows the results obtained when applying Benders decomposition (see Section 5) to two different case studies drawn from the rapid transit network "Cercanías Madrid" operated by RENFE: first line C5 and second lines C3 and C4.

Line C 5: Fig. 5(a) shows the evolution of the Benders upper and lower bounds for the line C5 case study. We have set up a maximum computational time of 14,400 s. As we can see, Benders decomposition fails in finding the optimal solution within the maximum time limit: the optimality gap provided by this approach is around 83%.

Fig. 4. Line C4.

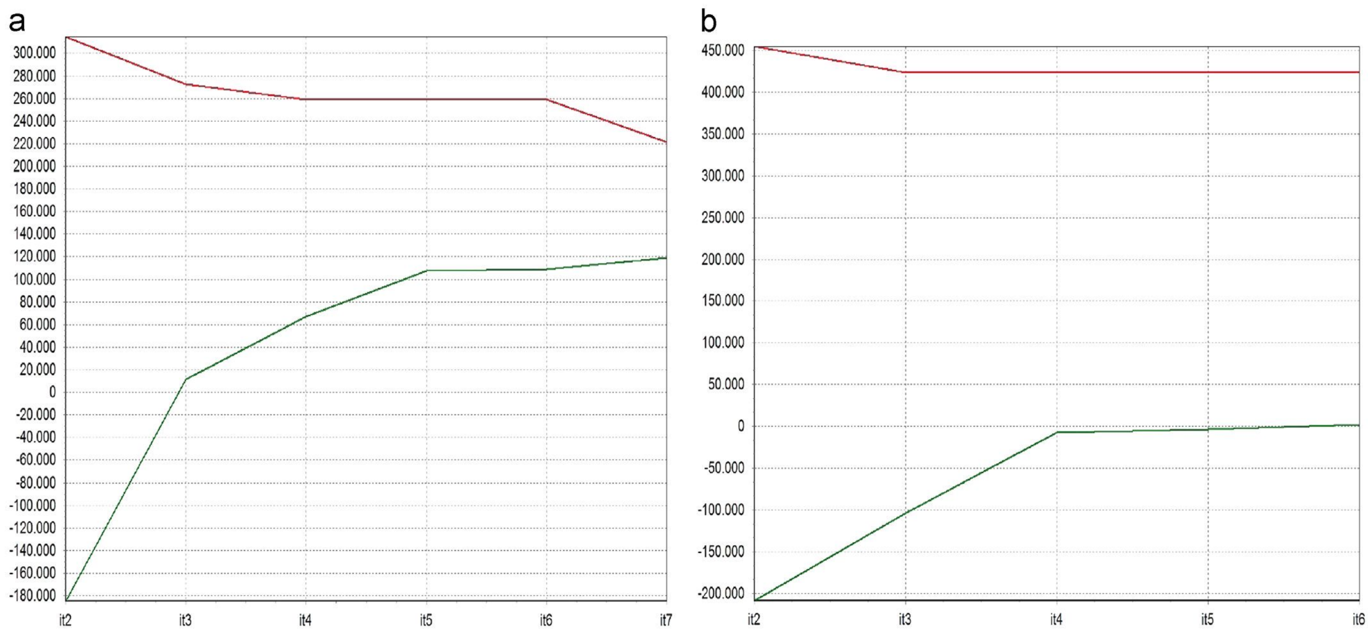

Fig. 5. Evolution of the Benders upper and lower bounds. (a) Line C5 case study. (b) Lines C3 and C4 case study.

Lines C 3 *and C* 4: Similarly, Fig. 5(b) shows the evolution of the Benders upper and lower bounds for the lines C3 and C4 case study. We have set up the same maximum computational time as before. Again, Benders decomposition fails in finding the optimal solution within the maximum time limit: the optimality gap provided is around 4150%.

Applying Benders decomposition with $AOBC_{it}$, we have observed that the algorithm fails in providing acceptable solutions in a reasonable number of iterations or time, namely computational effort. During the initial iterations it provides solutions with an unaffordable number of composition changes and empty services. Therefore, the algorithm runs many iterations without providing any improvement with respect to sequentially obtained solutions, making the Benders *MM* bigger and more difficult to solve due to the large number of Benders optimality cuts. We develop a heuristic in the next subsection so as to solve the problem within reasonable computational time and to improve sequentially obtained solutions.

6.2. *Polishing sequentially obtained solutions*

The proposed Benders *MM* is composed of (30), (3)–(16) and optimality cuts (31). In these cuts, two different terms are appearing on the right hand side: the first one representing ending operations and the second one beginning operations. The issue is to design a *smart schedule* to achieve robust sequences. However, the schedule (i.e., the timetable) for a train service $\ell \in L^t$ is already fixed. Hence, it makes no sense to try to change it. Nevertheless, the schedule for empty services and shunting operations is not fixed, we may decide on it. Therefore, we will only include those terms referring to empty trains and shunting operations in the optimality cuts.

Consequently, a new set I_1 is defined. It is a set of operations composed of empty trains, aggregations and disaggregations. So, the new Benders based Heuristic $AOBC_{it}$ ($HAOBC_{it}$) are showed in (32):

$$\upsilon \geq \sum_{i \in I_1}\sum_{s \in Sc}\sum_{c \in C}\tilde{\alpha}_{i,c}\phi^{c}_{i,s}\kappa^{c,it}_{i,s} + \sum_{j \in I_1}\sum_{s \in Sc}\sum_{c \in C}\tilde{\beta}_{j,c}\varphi^{c}_{j,s}\lambda^{c,it}_{j,s} \quad \forall it \in HAOBC_{it} \tag{32}$$

The proposed heuristic is not an exact method: we are not obtaining exact solutions to the integrated model but improving sequentially obtained results. Therefore, a stopping criterion must be selected. For these case studies this criterion will be as follows: stop if a specified number of iterations has elapsed in total or since the last Best Solution was found. Then, the heuristic will terminate if five consecutive iterations do not improve the Best Solution found. We define the Best Solution as that one with the lowest expected propagated delay.

Consequently, the Benders based heuristic consists of obtaining an initial RS assignment. Then, the train routing is obtained, and based on that routing a new RS assignment is obtained. Obviously, the RS assignment and shunting schedules change every iteration. For the first iteration we provide to the Benders based heuristic with the rolling stock assignment problem solution.

Line C 5: Table 1 shows the evolution of the solution for each of the performed iterations for Line C5. In the first column the iteration number is shown; the number of train units used in the proposed solution (#C) in the second column; the train service operating costs (TSOC) in the third column; the empty service operating costs (ESOC) in the fourth column; the unattended passenger costs (UPC) are shown in the fifth column; the number of composition changes (#CC) in the sixth column; the expected delay propagation (EDP) in minutes in the seventh column; and the solution time (ST) in seconds in the last column.

At each iteration a different solution is obtained. It may seem that some solutions are equal in some costs. However, we must account for the fact that the empty trains and shunting schedule is being changed. Consequently, operating costs might be equal but the sequences are being changed and so does the EDP. Moreover, each iteration provides the operator with a feasible planning which may be implemented. The operator may decide to choose among them according to several criteria such as train units maintenance and crew scheduling. Therefore, this approach provides different solutions to the operator.

We compare the obtained heuristic integrated solution, the best solution found by the Benders decomposition (with a computational time limit of 14,400 s), the sequentially obtained solution and the current solution in Table 2. In the second column #C is shown. TSOC are in the third column, ESOC in the fourth one, UPC in the fifth column, #CC in the sixth one, EDP in the seventh column, and the expected delay reduction (EDR) in the last one. The Benders based heuristic chooses the solution that produces the lowest EDP. We can see how the obtained solution in the heuristic integrated approach (Benders Heuristic) improves the EDP with respect to the rest of solutions: the solution obtained by the Benders decomposition (which is impractical due to the huge ESOC and #CC), the current solution operated by RENFE (Current solution) and the one obtained by the sequential approach (RS and TR solution, see [7] and [8]). In addition, the number of performed composition changes is reduced with respect to the solution obtained by the sequential approach; this reduction is always good news because of the possibility of malfunction: the lower the number of composition changes, the greater of the robustness degree. However, the achievement of this robustness is not for free. We acknowledge that TSOC and ESOC increase a bit in the heuristic integrated solution as compared to the sequential approach. Moreover, UPC is also increased with respect to the current solution. Nevertheless, this might be seen as a dummy increase: the network operator considers service to be of good quality when standing passengers do not exceed a density of 3.5 pax/m^2. We have therefore attempted to match this objective: we have imposed the upper limit to passenger capacity to that one corresponding to a density of 4 pax/m^2 (so as to have an average density which is lower). Thus we get a service quality roughly equal to the one the operator is looking for. However, this does not mean that the remaining passengers are unattended, but they will be able to board the trains in the real life (the real capacity is not the one corresponding to a density of 4 pax/m^2). In the current solution UPC is much lower because all the train services have the greatest composition assigned. For more details on this see [8]. However, a great EDR is obtained. In order to get a deeper insight in the expected delay reduction, the distribution of the propagated delay is shown in Fig. 6 for the integrated and sequential approaches. In the integrated approach the number of operations connections (defined by two different and consecutive operations performed by the same train unit) (vertical axis) with positive expected propagated delay (horizontal axis, expressed in minutes) is slightly reduced. Consequently, the number of needed swapping operations will be also reduced. Therefore, the integrated planning turns out to be more robust and smoother to be operated.

In the integrated approach the computational time is of 1620 s. It is higher than in the sequential approach. However, the obtained solution has fewer composition changes than the solution in the sequential approach and has fewer operation connections with positive expected propagated delay: the solution is more robust. Furthermore, the are fewer operation connections with positive expected propagated delay. In addition, computational time is still reasonable for the planning horizon we are working on. Altogether, the integrated approach becomes an interesting way of tackling the problem of the circulations for the operator.

Lines C 3 and C 4: Again, we use the proposed Benders heuristic to solve the integrated model. In Table 3 we can see the evolution of the solution for each of the performed iterations. The results obtained in this case study lead us to similar conclusions as the ones presented for the line C5 case study. In the first column the iteration number is shown. The #C in the second column, the TSOC in the third column, the ESOC in the fourth column, the UPC are shown in the fifth column, the #CC in the sixth column, the EDP in minutes in the seventh column, and the ST in the last column. In each iteration a different solution is obtained. Among all the feasible solutions provided by the heuristic we choose the solution with the lowest EDP.

We compare the obtained heuristic integrated solution, the best solution found by the Benders decomposition (with a computational time limit of 14,400 s), the sequentially obtained solution and the current solution in Table 4. Table 4 may be read as Table 2. We can see how the obtained solution by the heuristic integrated approach improves the EDP with respect to the rest of the solutions and reduces the number of performed composition changes with respect to the solution obtained by the sequential approach. Note that the solution found by the Benders decomposition is impractical due to the huge number of ESOC and #CC.

Table 1
Line C5 Benders heuristic solution at each iteration.

Iteration	#C	TSOC	ESOC	UPC	#CC	EDP	ST
1	64	80,099.7	1265	3554	20	111.8	28.68
2	65	80,563.44	1645.3	3248	14	72.05	54.07
3	65	80,570.64	1393.04	3248	16	75.01	46.16
4	65	80,570.64	1490.24	3248	14	67.16	187.39
5	66	80,727.6	1359.04	3248	14	68.69	60.92
6	66	80,346	1553.92	3639	10	69.76	258.17
7	66	80,413.68	1457.44	3554	14	61.03	55.18
8	66	80,346	1526.32	3639	12	71.89	349.83
9	66	80,346	1563.12	3639	10	69.54	147.44
10	66	80,360.4	1475.2	3696	12	66.07	261.13
11	66	80,346	1574.32	3639	10	63.53	112.96
12	66	80,742	1612.1	3326	12	70.48	58.07

Table 2
Line C5: Benders heuristic integrated solution, Benders decomposition solution, sequential (RS and TR) solution and current solution.

Solutions	#C	TSOC	ESOC	UPC	#CC	EDP	EDR (%)
Benders heuristic	66	80,413.68	1457.44	3554	14	61.03	56.9
Benders decomposition	70	84,321.96	19,852.22	3222	45	122.36	13.44
RS and TR	64	80,099.7	1265	3554	20	111.80	20.91
Current	74	109,765.2	2232.1	874	0	141.37	-

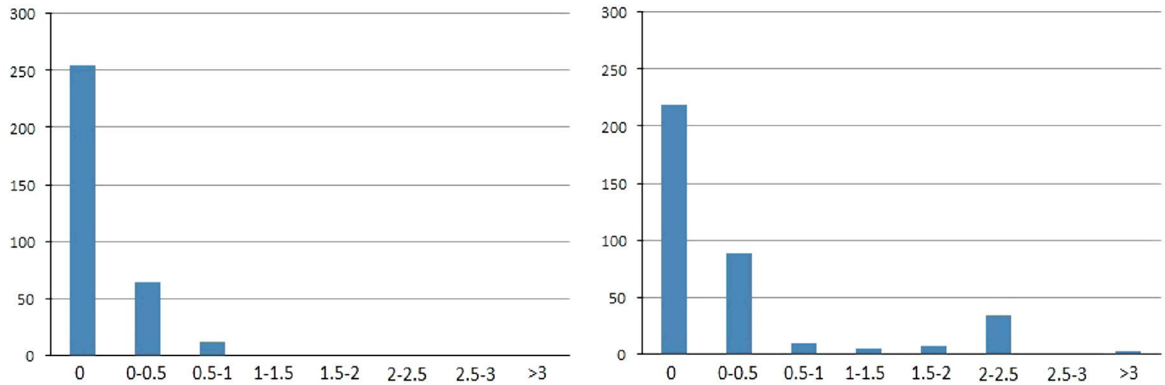

Fig. 6. Line C5: number of operations connections for each propagated delay in the integrated approach (left side) and in the sequential approach (right side).

Table 3
Lines C3 and C4 Benders based heuristic solution at each iteration.

Iteration	#C	TSOC	ESOC	UPC	#CC	EDP	ST
1	63	87,837.52	4090.83	2177	34	163.98	132.05
2	62	88,937.23	3810.84	2177	38	179.2	213.55
3	64	89,811.43	4624.2	3403	32	152.25	350.08
4	64	88,653.73	3796.80	2723	32	147.95	581.20
5	66	90,206.71	4179	2177	36	164.46	550.32
6	65	90,532.66	3780.6	2077	40	183	684.89
7	65	90,128.2	4061.61	2177	38	182.37	487.13
8	66	90,754.33	3608.19	2177	38	182.37	605.78

Again, in order to get a deeper insight in the EDR, the distribution of the propagated delay is shown in Fig. 7 for the integrated and sequential approaches. In the integrated approach the number of operations connections (vertical axis) with positive expected propagated delay (horizontal axis, expressed in minutes) is reduced. Consequently, the number of needed swapping operations will be also reduced.

The computational time is increased compared to the line C5. Now, the time needed to solve the model is of 3605 s. This is due to the fact that in this case two different lines are being solved, there are more depot stations and possibilities for shunting operations are greater. Thus, Benders heuristic's cuts are larger and the problem becomes more difficult to be solved. Nevertheless, as the purpose of this work is to propose the schedule well in advance to the day of operations execution, the computational time required is still reasonable.

6.3. Sensitivity analysis and multiobjective optimization

This subsection is devoted to the objective function: first, we conduct several sensitivity analysis on various objective function parameters (we assume that $\omega=0.5$); second, we perform a multiobjective optimization study (i.e., we vary the ω values).

Sensitivity analysis: We acknowledge that the presented objective function combines actual operating costs with other terms that are related to penalties in order to introduce robustness to the system ($km_\ell, upc_{a,\ell}, \vartheta_s$) and costs per delay and need for human resources (ψ,ζ), which are difficult to be estimated. Therefore, it is very important to investigate the sensitivity of our results to changes in these parameter values. The results of the sensitivity analysis to these model parameter values for each of the studied lines are presented in Tables 5 and 6. The parameters are varied within -20% to $+20\%$ of their nominal values. The first column of the tables shows the parameter to be varied (the percentage variation is shown in the first row). Each element in the tables presents the percentage variation in the objective function with respect to the objective function value for the nominal parameter value (i.e., percentage variation of 0). As shown in the tables, the objective function values vary between -14.32% and 12.38% and are reasonably stable to significant variations in model parameters.

Multiobjective optimization: The objective function in (1) accounts for two different main costs: RS assignment costs and TR costs. Consequently, giving different weights to each of the costs by means of ω yields different solutions. The operator might be interested in knowing these different solutions because we are making decisions in the presence of trade-offs between two conflicting objectives: minimizing rolling stock assignment costs means that we incur into greater train routing costs and vice versa (see Fig. 8). Fig. 8 shows the pareto-optimal fronts for the line C5 and the lines C3 and C4. Because our model is a mixed-integer programming model, each different value for ω does not necessarily provide a different solution. Due to this reason we use discontinuous lines to depict the pareto-optimal fronts. Each cross (for the line C5) and dot (for the lines C3 and C4) represent a solution for a ω value.

6.4. Summary

The proposed algorithmic framework allows us to find solutions to the circulation problem in an integrated fashion. This integration leads us to a huge optimization model, thus we define a heuristic to solve it. The solutions we have obtained with the integrated approach are better than those ones obtained with the sequential approach. We have presented two case studies, which are drawn from line C5 and lines C3 and C4 in RENFE's regional network in Madrid:

- *Line C 5*: Applying a sequential approach we were able to reduce the expected delay by a 20.91% compared to the current solution operated by RENFE (see [7]). Now, applying the integrated approach we polish sequentially obtained solutions and we show that a reduction of a 56.9% is possible.
- *Lines C 3 and C 4*: Again, applying a sequential approach we were able to reduce the expected delay by a 16.19% (see [7]). However, applying the integrated approach we polish sequentially obtained solutions and we show that a reduction of a 24.4% is possible.

Therefore, solutions from the integrated approach clearly outperform solutions from the sequential approach. Moreover, we also state that this great reduction in expected delays does not mean an unaffordable increase in operating costs. The better feedback we have by solving the two problems in an integrated way makes this improvement possible.

Table 4
Lines C3 and C4: Benders heuristic integrated solution, Benders decomposition solution, sequential (RS and TR) solution and current solution.

Solutions	#C	TSOC	ESOC	UPC	#CC	EDP	EDR (%)
Benders heuristic	64	88,653.73	3796.80	2723	32	147.959	24.4
Benders decomposition	68	95,621.39	21,254.47	1896	88	178.36	8.83
RS and TR	63	87,837.52	4090.83	2177	34	163.983	16.19
Current	72	136,633.86	6083.88	1550	0	195.65	–

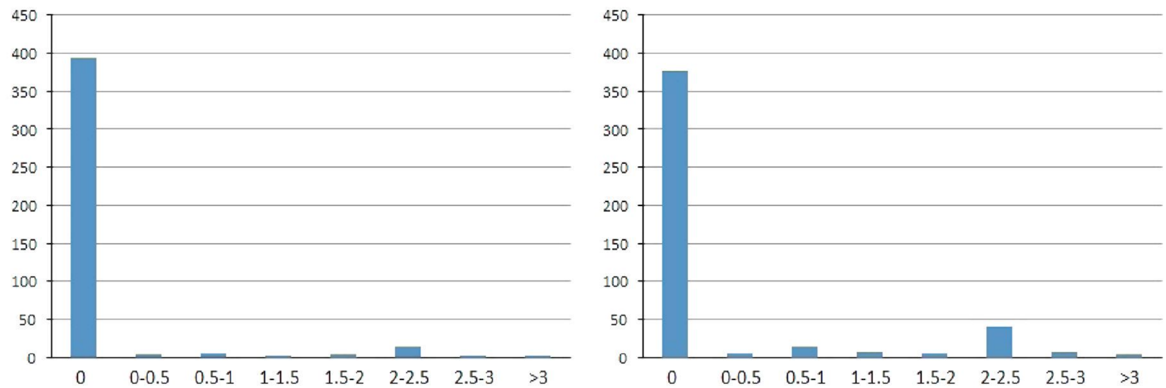

Fig. 7. Lines C3 and C4: number of operations connections for each propagated delay in the integrated approach (left side) and in the sequential approach (right side).

Table 5
Sensitivity analysis for line C5.

Parameter	–20%	–15%	–10%	–5%	0%	5%	10%	15%	20%
km_{ℓ}	1.80	3.34	1.30	1.41	0	–0.06	0.97	0.14	0.78
$upc_{a,\ell}$	1.42	4.78	8.80	5.25	0	1.36	2.61	0.25	0.96
ϑ_s	–2.56	–1.41	0.37	–0.14	0	0.73	1.87	2.04	2.66
ψ	–1.05	1.33	0.48	–0.64	0	2.77	2.04	–0.19	7.33
ζ	–10.23	–5.25	–3.27	–2.17	0	5.30	9.51	9.61	12.38

Table 6
Sensitivity analysis for lines C3 and C4.

Parameter	–20%	–15%	–10%	–5%	0%	5%	10%	15%	20%
km_{ℓ}	1.41	2.04	1.56	1.01	0	–0.86	1.87	2.48	2.81
$upc_{a,\ell}$	–3.23	–1.81	1.01	2.59	0	0.69	1.19	2.51	3.61
ϑ_s	–1.66	–1.19	0.78	–0.44	0	0.39	2.78	3.12	3.96
ψ	–2.50	–1.01	0.13	0.41	0	1.42	1.97	2.19	5.43
ζ	–14.32	–7.21	–4.74	–2.87	0	3.03	5.17	8.51	10.81

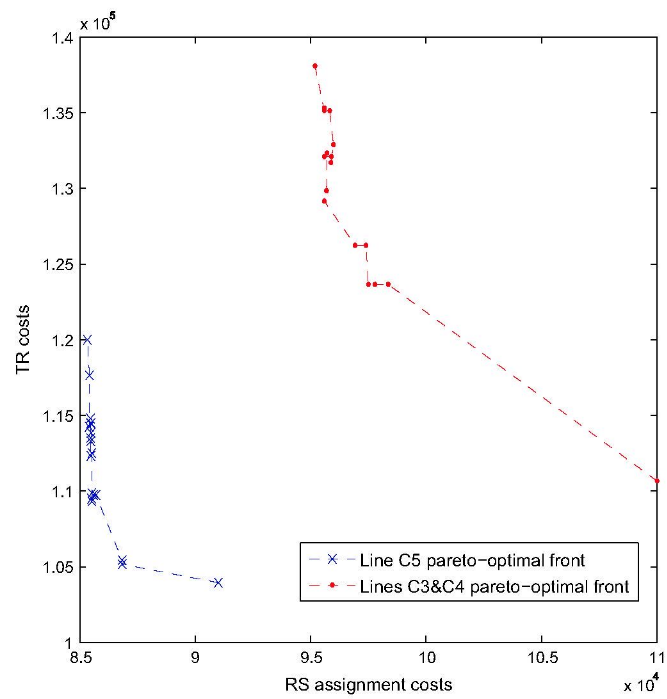

Fig. 8. Pareto-optimal fronts for lines C5 and C3 and C4.

7. Conclusions

We have proposed a new approach to solve the rolling stock circulation problem. We address two different problems: the rolling stock assignment problem and the train routing problem. Up to now, they have been solved in a sequential manner, that is, the solution of one of them is known before solving the other one. Our approach proposes to integrate both problems. Consequently, it overcomes the drawbacks of the sequential approach, namely its iterative nature and the likely of obtaining sub-optimal or even infeasible solutions.

The integrated approach is a good frame to improve the robustness degree of the system. We get the robustness through different points of view: rolling stock operations and operations connections. Among rolling stock operations we may cite empty trains and composition changes. They are always difficult operations and complicate the network performance. A way of introducing robustness here is penalizing selectively these operations in order to ameliorate their possible negative effects in the network operation. The other way of introducing robustness we have presented is related to schedule performance and punctuality. In order to perform operations connections properly, the schedule must be matched. However, the schedule is not always operated as planned and deviations from the planned operations may produce delays in operations. We account for these possible delays in order to produce a resistant schedule to them.

The proposed integrated model to solve the rolling stock circulations has an enormous size to be solved by the current commercial software. Therefore, we propose to decompose it using Benders decomposition. Using this technique the submodel may be reformulated in order to make it easier to be solved.

However, for computational reasons we propose a Benders based heuristic to solve the proposed model.

Computational experiments show how the current solution operated by RENFE can be improved: more robust and smoother solutions are obtained. RENFE planners do not use operations research techniques for planning purposes. Therefore, the planning is greatly improved. Furthermore, the proposed integrated approach also outperforms the solution obtained using operations research techniques in a sequential manner. We are able to produce plans with fewer composition changes which are considered to be dangerous by planners. Selectively penalizing them we ameliorate their probability to produce negative cascading effects in the network. In addition, we also account for delays: we reduce the number of operations connections with positive propagated delay, thus reducing the number of re-scheduling operations (i.e., swapping operations) and making it easier for the operator to recover.

The solution provided by our integrated approach is the schedule to be implemented by the operator. Consequently, the available planning horizon is enough to consider the needed computational times reasonable for implementation. Nevertheless, development of more intelligent heuristics to reduce the model size and computation costs may be defined.

Acknowledgments

This research was supported by project grant TRA2011-27791-C03-01 by the "Ministerio de Economía y Competitividad, Spain". We also want to thank the anonymous referees for their helpful comments on the paper.

References

[1] Alfieri A, Groot R, Kroon LG, Schrijver A. Efficient circulation of railway rolling stock. Transp Sci 2006;40:378–91.
[2] Almeida D, Marcos N, Vianey S, Weber C. Robustness of rolling stock schedules. Arrival-TR-0129. Available from ⟨http://arrival.cti.gr/uploads/Documents.0129/ARRIVAL-TR-0129.pdf⟩ [accessed September 2013].
[4] Barnhart C. Airline schedule optimization. In: Belobaba P, Odoni A, Barnhart C, editors. The global airline industry. United Kingdom: Wiley; 2009 ISBN: 978-0-470-74077-4.
[5] Cacchiani V, Caprara A, Galli L, Kroon LG, Maróti G, Toth P. Railway rolling stock planning: robustness against large disruptions. Transp Sci 2012;46:217–32.
[6] Cadarso L. Robustness and recoverability in transport logistics [Ph.D. thesis]. Technical University of Madrid; 2013.
[7] Cadarso L, Marín A. Robust routing of rapid transit rolling stock. Public Transp 2010;2:51–68.
[8] Cadarso L, Marín A. Robust rolling stock in rapid transit networks. Comput Oper Res 2011;32:1131–42.
[9] Cadarso L, Marín A. Integrated robust airline schedule development. Procedia – Soc Behav Sci 2011;20:1041–50.
[10] Cadarso L, Marín A. Integration of timetable planning and rolling stock in rapid transit networks. Ann Oper Res 2012;199(1):113–35.
[11] Cadarso L, Marín A. Robust passenger oriented timetable and fleet assignment integration in airline planning. J Air Transp Manag 2013;26:44–9.
[12] Cadarso L, Marín A, Maróti G. Recovery of disruptions in rapid transit networks. Transp Res E-Log 2013;53:15–33.
[13] Cadarso L, Vaze V, Barnhart C, Marín A. Integrated airline scheduling: considering competition effects and the entry of the high speed rail, Transp Sci 2014, under review.
[14] Campos J, De Rus G. Some stylized facts about high-speed rail: a review of HSR experiences around the world. Transp Policy 2009;16(1):19–28.
[16] Fioole PJ, Kroon LG, Maróti G, Schrijver A. A rolling stock circulation model for combining and splitting of passenger trains. Eur J Oper Res 2006;174(2):1281–97.
[17] Ghoseiri K, Szidarovszky F, Asgharpour MJ. A multi-objective train scheduling model and solution. Transport Res B – Methods 2004;38(10):927–52.
[18] Huisman D, Kroon LG, Lentink RM, Vromans MJCM. Operations research in passenger railway transportation. Stat Neerl 2005;59(4):467–97.
[19] Laporte G, Mesa JA, Ortega FA. Optimization methods for the planning of rapid transit systems. Eur J Oper Res 2000;122(1):1–10.
[20] Lohatepanont M, Barnhart C. Airline schedule planning: integrated models and algorithms for schedule design and fleet assignment. Transp Sci 2004;38:19–32.
[21] López F, Codina E, Marín A. Integrating network design and line frequency setting in rapid transit systems. Public Trans 2014, under review.
[22] Marín A, Mesa JA, Perea F. Integrating robust railway network design and line planning under failures. In: Ahuja RK, et al. (Eds.), Robust and online large-scale optimization, lecture notes in computer science, vol. 5868; 2009. p. 273–92.
[23] Maróti G. Operations research models for railway rolling stock planning [Ph.D. thesis]. Eindhoven, Technische Universiteit Eindhoven; 2006.
[24] Mercier A, Cordeau JF, Soumis F. A computational study of Benders decomposition for the integrated aircraft routing and crew scheduling problem. Comput Oper Res 2005;32(6):1451–76.
[25] Papadakos N. Integrated airline scheduling. Comput Oper Res 2009;36(1):176–95.
[26] Peeters M, Kroon LG. Circulation of railway rolling stock: a branch-and-price approach. Comput Oper Res 2008;35:538–56.
[27] Pita J, Barnhart C, Antunes AP. Integrated flight scheduling and fleet assignment under airport congestion. Transp Sci: 1–16. ⟨http://dx.doi.org/10.1287/trsc.1120.0442⟩.
[28] Repolho HM, Antunes AP, Church RL. Optimal location of railway stations: the lisbon-porto high-speed rail line. Transp Sci 2013;47(3):330–43.
[29] Walker CG, Snowdon JN, Ryan DN. Simultaneous disruption recovery of a train timetable and crew roster in real time. Comput Oper Res 2005;32:2077–94.
[30] Wolsey LA. Integer programming, Wiley interscience series in discrete mathematics and optimization; 1998. ISBN: 978-0-471-28366-9.