Train dispatching

Leonardo Lamorgese, Carlo Mannino, Dario Pacciarelli, Johanna Törnquist Krasemann

Citation:
Lamorgese L., Mannino C., Pacciarelli D., Krasemann J.T. (2018) Train Dispatching. In: Borndörfer R., Klug T., Lamorgese L., Mannino C., Reuther M., Schlechte T. (eds) Handbook of Optimization in the Railway Industry. International Series in Operations Research & Management Science, vol 268. Springer, Cham

Train dispatching - uncorrected proof

Leonardo Lamorgese * Carlo Mannino * Dario Pacciarelli †
Johanna Törnquist Krasemann

Abstract

The challenge of managing disturbances and delays in railway traffic systems has received significant attention in the operations research community during the past 20-30 years. It is a complex problem with many aspects and constraints to consider. This chapter defines the problem and summarizes the variety of model types and solution approaches developed over the years, in order to address and solve the train dispatching problem from the infrastructure manager perspective.

Despite all the research efforts, it is, however, only very recently that the railway industry has made significant attempts to explore the large potential in using optimization-based decision-support to facilitate railway traffic disturbance management. This chapter reviews state-of-practice and provides a discussion about the observed slow progress in the application of optimization-based methods in practice. A few successful implementations have been identified, but their performance as well as the lessons learned from the development and implementation of those system are unfortunately only partly available to the research community, or potential industry users.

1 Introduction

1.1 Background and scope

The challenge of managing disturbances and delays in railway traffic systems has most likely existed ever since the launch of the first public railway system in the beginning of the 19th century. With enlargements of the railway systems and service networks, as well as introduction of new technology and more complex organizational structures, the potential sources of faults and their knock-on effects increase. Managing railway traffic networks is today not only a technical achievement and challenge, but appears at times to become more of an organizational challenge. An organizational challenge

*SINTEF ICT Applied Mathematics, Oslo, e-mail: leonardo.lamorgese@sintef.no, carlo.mannino@sintef.no

†Università degli Studi Roma Tre, Dipartimento di Ingegneria, Rome (Italy), e-mail: pacciarelli@ing.uniroma3.it

in the sense that there are nowadays often so many sub-systems, stakeholders and dependencies which hampers the ability to overview and manage network activities in a proactive way and with a system perspective. Railway system stakeholders thus need to jointly and stepwise start incorporating effective decision-supporting protocols and software in a much larger extent than now and evaluate the strengths and weaknesses in a systematic and transparent manner so that lessons learned reach beyond individual project groups and system suppliers.

The efforts dedicated in industry and in the research community to develop and evaluate principles, methods and software for decision-support in railway traffic management have increased significantly over the past 20 years. Larger European projects such as COMBINE, ARRIVAL and ON-TIME has resulted in relevant results and pinpointed several challenges, but more emphasize on practical applications and evaluations is needed. The results from the large national tenders seen in Europe recent years, and the forthcoming developed large-scale traffic management systems, will hopefully contribute to an increased knowhow in the field and shed some light on many important practical and theoretical questions. We will discuss this aspect further later on in this chapter.

In this chapter, we first briefly introduce the research domain and main terminology with emphasize on optimization-based decision-support for railway traffic management during disturbances. In section 2, two common alternative modeling approaches and problem formulations are presented and discussed. In section 3, common types of algorithmic approaches are presented and discussed, while section 4 presents current and planned practical implementations. Section 5 presents some conclusions and pointers for future work.

1.2 Aspects of disturbance management in railway traffic systems

A disturbance in a railway network can occur due to a smaller incident such as an over-crowded platform and unexpectedly long boarding times causing minor delays, which the affected train may be able to recover from if there is sufficient buffer in the timetable. Disturbances can also be more significant and occur due to e.g. rolling-stock breakdowns, power shortages, or signaling system failures.

Larger disturbances are in the context of railway traffic management sometimes referred to as disruptions, although the words generally can be considered synonymous. The Oxford online dictionary defines disruption as *disturbance or problems which interrupt an event, activity, or process.*

The distinction between smaller and larger disturbances has been discussed in e.g. [3]. There, the following definition is used: *...disturbances are relatively small perturbations of the railway system that can be handled by modifying the timetable, but without modifying the duties for rolling stock and crew. Disruptions are relatively large incidents, requiring both the timetable and the duties for rolling-stock and crew to be*

modified (ibid). Hence, the distinction primarily is based on what type of actions that may be needed to cope with the incident rather than the initial sources of disruption. In this chapter we adopt the same distinction for sake of clarity.

Disruptions often result in significant knock-on delays and longer period of partial system unavailability. The railway system state transition over time can be illustrated as a bathtub, see e.g. [14], where the traffic is reduced to function only partially during the disruption. That reduced level of traffic is then maintained until the system goes back to full capacity again via a transition plan.

When a railway traffic network suffers from a disturbance or disruption, which affects the scheduled railway transport services, the timetable needs to be modified. The re-scheduling of the timetable consists of two main parts:

1. Traffic re-scheduling, where focus is on network capacity and the need of the infrastructure manager (IM) to revise the timetable and allocation of track resources for the affected trains to minimize delays;

2. Transport service re-scheduling where focus is on the transport operating companies (TOC) and their need to handle the timetable from a train service point of view explicitly considering train connections and the effects on the rolling-stock and crew schedules.

The latter part includes the delay management problem, where emphasize is on effective policies for managing trains connections and passenger flows during disturbances, in order to minimize passenger delays given a predefined set of available train services. In contrast to traffic re-scheduling, the delay management problem does traditionally not consider network capacity issues although the recent trend is to incorporate an increasing level of detail and realism in the models [11].

Although the majority of research so far has focused on the mentioned perspectives and types of re-scheduling problem individually, the interest in integrated approaches is increasing, see e.g.[7].

Depending on the organizational structure of the railway systems and what authority and control the traffic managers have, the decision-making may be distributed between several different stakeholders. In fully deregulated networks such as the national railway systems of Sweden and Norway, the control of the infrastructure and traffic management lies on a neutral national transport authority, while the trains and associated transport services are operated by several different private companies. The decision-making during disturbances and disruptions is then depending on two, or more, different organizations. The different types of re-scheduling decisions can be divided as follows [15]:

- Re-timing of trains by allocating new arrival and departures times, including modification of speed profiles and halting schedules.

- Re-ordering of trains by adjusting the meet-pass plans.

- Local re-routing, by allocating alternative tracks on the line between two stations, or within the stations.
- Global re-routing by allocating alternative paths in the network.
- Cancellations and/or turning trains earlier than expected.

The first three can normally be made by the IM without consulting the TOCs, but the last two requires consultation with affected TOCs. In this chapter, we will continue discussing only real-time railway traffic disturbance management and focus on the infrastructure manager perspective during re-scheduling.

The development and application of such computational real-time re-scheduling support encompasses several challenges related to:

1. Human-computer interaction and requirements engineering concerning how to define and configure the computational support as functionalities.
2. Specification and formulation of the specified re-scheduling problems as well as development of appropriate solution methods.
3. System integration and communication including input data availability.

Here we focus on the second aspect and particularly optimization-based models and solution approaches. The research concerning development and application of decision-support for railway traffic disturbance management has received significant attention, which can be seen by comparison of different surveys and literature reviews over the years, see e.g [34], [20] and [13].

Proposed models and problem formulations can be described and compared with respect to a number of key properties such as infrastructure representation and level of granularity, time representation and objective(s). Railway traffic and the associated train occupation of network resources is often modelled as train events, where the events are assigned specific time slots for the associated network resources. The problem of deciding a) which resource to assign to each train event, b) in what order different train events should be allocated the resources and c) during which time period, is then the core problem. This problem is often referred to as the Train Dispatching problem (TD), see [17] and can be seen as a Job Shop Scheduling Problem with *blocking and no-wait constraints* [26].

The TD problem is formulated in several different ways depending on how capacity limitations of the network are modelled. The models are often classified as macro models or micro models, although there is no exact definition of what details each level includes. Macro models typically disregards the railway signalling system and consider the lines between stations as a set of parallel tracks. Micro models consider these lines as a set of train paths, each defined by a chain of block sections, crossings and signals.

In [23], two alternative MILP formulations to model the capacity restrictions on stations on a macro level are proposed and benchmarked: A *non-compact* formulation

which counts the number of pairs of trains that simultaneously meet at a specific station and then ensures this respects the capacity limit, and a *compact* formulation where the station track occupancy by each train is modeled explicitly. Such approaches are often sufficient for single- and double-track lines, where there are less complex junctions and stations. Examples of models which focuses on complex and busy stations and network areas are presented in [6] and [29]. There are also hybrid models which combine the use of macro and micro models. For example, [5] use a macro model for the compensation zones of the network, with simple topology and low traffic, and a more detailed model for condensation zones, where it is necessary to include more detail in order to ensure feasible solutions. A different approach is proposed in [17], in which two-level strategies are adopted to optimally solve the dispatching problem on a large network, where macro models are used for the upper level and micro models are used for the lower level.

The majority of the problem formulations seen in the literature use a continuous time representation, where disjunctive constraints (also referred to as big-M constraints) are used to decide on the pairwise order of trains on allocated track segments. There are also a significant number of researchers who use a discrete time representation, see e.g. [4] and [27]. In [16] results from a comparative analysis of these two alternative ways of representing track occupancy over time are presented.

The objective(s) of the re-scheduling approaches have traditionally been to minimize train delays in different ways considering e.g. maximum consecutive delay, or train delays with different weights. The focus on minimization of passenger delays and inconvenience rather than train delays has, however, increased, see e.g. [33]. There is also a significant stream of research that also include aspects of energy-efficient driving. Such approaches consequently include a more fine-grained model of the infrastructure properties, network topology and train speed profiles in order to compute train trajectories instead of approximations of run times.

For a more detailed description of common alternative modelling approaches and problem formulations we refer to [21].

2 Alternative Graphs and Alternative Problem Formulations

This section focuses on microscopic models for train dispatching from the IM perspective, i.e., with the purpose of adjusting the timetable in response to disturbances in order to minimize train delays.

Train traffic is controlled worldwide by signals, interlocking and Automatic Train Protection (ATP) systems, which set up train routes, enforce train speed restrictions and impose a minimum safety separation between trains to avoid train accidents. Fixed block ATP systems ensure safety by allowing at most one train at a time running on a *resource* of the network. Examples of resources are the platforms of a station or the *block sections* of a line, i.e., portions of track between two consecutive signals. The

path of a train from its origin to its destination is therefore a sequence of resources, each assigned exclusively to that train for a specific time interval, called *occupation time*. This time takes into account the dwell time on a stop platform, or the minimum traversing time of a block section, which depends on several factors, such as the length of the block section, the train speed and length, and includes the time between the entrance of the train head (its first axle) in the block section and the exit of its tail (the last axle), plus additional time margins to release the occupied block section and to take into account the sighting distance.

Due to the safety systems, a *primary delay* of a train due to an unexpected delay may easily propagate to other trains in the network, causing *secondary delays*. In fact, at least in areas with dense traffic, the amount of secondary delay due to propagation may significantly exceed that of the originating primary delay (Goverde PT). The task of the dispatcher is then to adjust the timetable to limit as much as possible delay propagation by keeping a feasible plan of operations. Specifically, given a railway network and a set of train circulating in the network in a given time horizon, the TD problem is the real-time problem faced by the dispatchers, which consists in choosing:

1. a *route* for each train, i.e., a sequence of resources from its initial position, or entry point, to its final position, or exit point, such that the train can physically move from each resource to the next in the sequence;

2. a *sequence* of trains, for each resource traversed by multiple trains;

3. a *schedule* for each train, prescribing the time at which the train should start the occupation of each resource along its route from its entry to its exit point. The schedule is feasible if there is no *conflict*, i.e., if no two trains try to occupy the same resource at the same time. A train keeps a resource occupied from the start of the occupation for at least the prescribed occupation time. However, having reached the end of the resource, the train can keep it occupied for an additional time if the subsequent resource on its route is not available, and blocks the current resource preventing other trains from entering it. This fact may lead to a em deadlock status when there is a circular precedence among a set of trains, each waiting the resource blocked by another train in the set.

The main objective of the dispatcher is to design a deadlock-free schedule minimizing a function of train delays. Typical objectives proposed in the literature ranges from the minimization of the average delay, or of the maximum secondary delay [10], to the minimization of convex functions of the train delays [24]. Microscopic models to represent feasible solutions are based on the observation that by viewing the trains as jobs and the resources as machines, the train dispatching problem is similar to the job shop scheduling problem, in which the occupation time of a train on a resource corresponds to the processing time of a job on a machine. However, there are some specific aspects of the TD problem that must be taken into account. For example a train, having reached the end of a resource, cannot enter the subsequent one if the

latter is occupied by another train. In the scheduling theory this is known as a *blocking constraint*. Moreover, other constraints can be defined, such as a minimum departure time from specific resources (platform stops), or a maximum travelling time between pair of resources that can be useful to ensure that a train does not take too long to reach the next station. The latter constraint can be modelled as a generalized no-wait constraint. Hence, the TD problem can be viewed as a job shop scheduling problem with blocking and no-wait constraints. Once a routing is fixed for each train, this type of problem can be effectively formulated by the alternative graph formulation [26]. The alternative graph generalizes the disjunctive graph model of the job shop scheduling problem by Roy and Sussmann [30].

An alternative graph is a triple $\mathcal{G} = (N, F, A)$, with N being the set of nodes, F the set of *fixed* directed arcs and A the set of pairs of *alternative* directed arcs. Arcs in F and A are weighted, and let w_{ij} be the length of arc (i, j). A *selection* S is a set of arcs obtained from A by choosing at most one arc from each alternative pair. The selection is *complete* if exactly one arc from each pair in A is selected. Given a pair $[(i, j), (h, k)] \in A$ and a selection S, if $(i, j) \in S$ then arc (i, j) is *selected* and arc (h, k) is *forbidden*. The pair is *unselected* if neither (i, j) nor (h, k) is selected in S. Given a selection S, $\mathcal{G}(S)$ denotes the digraph $(N, F \cup S)$. The selection S is *consistent* if the graph $\mathcal{G}(S)$ has no positive length cycles. An *extension* S' of a consistent selection S is a consistent selection such that $S \subset S'$. Given two nodes $i \in N, j \in N$, then $l^S(i, j)$ denotes the length of the longest path from node i to node j in $\mathcal{G}(S)$.

In the alternative graph formulation of the TD problem, set N includes two dummy nodes 0 and n, called *start* and *finish* respectively, such that for each node $i \in N$ there is a directed path from 0 to i and a path from i to n in the digraph $\mathcal{G}(\emptyset) = (N, F)$. To each node in $i \in N \setminus \{0, n\}$ is associated the event that a train starts the occupation of a resource. Let τ_i and ρ_i be the train and the resource associated to event i, respectively. Finally, let t_i be a variable associated to the starting time of event $i \in N \setminus \{0\}$, while $t_0 = 0$. Note that setting a value t_i for all $i \in N \setminus \{0, n\}$ corresponds to choosing a schedule for all trains.

Each arc (i, j), either fixed or alternative, represents a precedence relation constraining the start time (t_j) with respect to the start time t_i. A fixed arc $(i, j) \in F$, where $\tau_i = \tau_j$ and ρ_j is the resource traversed by τ_i immediately after ρ_i, represents the constraint $t_j \geq t_i + w_{ij}$, where w_{ij} is the traversing time of resource ρ_i by τ_i. A fixed arc $(0, i) \in F$ represents the entrance of train τ_i in the network, or the departure time of τ_i from a platform stop ρ_i, with w_{0i} being the entrance time or the departure time of τ_i, respectively. A fixed arc $(i, 0) \in F$, with $w_{i0} < 0$, represents a firm deadline constraint $t_i \leq -w_{i0}$ for the start of occupation of train τ_i on resource ρ_i (recall that $t_0 = 0$). A deadline arc can be used, e.g., to represent the initial position of train τ_i at time 0, in combination with arc $(0, i)$ with $w_{0i} = -w_{i0}$.

A fixed arc $(i, n) \in F$, where i represents the arrival of τ_i at a platform stop ρ_i, is used to compute the delay of τ_i at a station. To this aim, let δ_i the arrival time of τ_i at ρ_i scheduled in the published timetable. By setting $w_{in} = -\delta_i$, arc (i, n) represents

the constraint $t_n \geq t_i - \delta_i$, i.e., the arrival delay of τ_i at ρ_i. Similarly, if ϵ_i denotes the originating delay of τ_i, setting $w_{in} = -\delta_i - \epsilon_i$ corresponds to the consecutive delay of τ_i at ρ_i. Hence, the minimization of the maximum (consecutive) delay can be obtained by minimization of t_n.

Alternative arcs are used to represent sequencing decisions. Given two trains τ_i and τ_h traversing the same resource $\rho_i = \rho_h$, is necessary to decide a sequencing for the two trains. Let ρ_j and ρ_k be the resources traversed by τ_i and τ_h immediately after ρ_i and ρ_h, respectively (see, Figure 2). Then, arcs (i, j) and (h, k) are added to F with weights w_{ij} and w_{hk} equal to the traversing time of ρ_i by τ_i and τ_h. In the example in Figure 2, two trains A and B share resources 1 and 4. Hence, for resource 1, $\tau_i = \tau_j = A$, $\tau_h = \tau_k = B$, $\rho_i = \rho_h = 1$, $\tau_j = 2$, $\tau_k = 3$. The pair $[(j, h), (k, i)] \in A$ represents the two sequencing alternatives. Arc (j, h) corresponds to giving precedence to train $\tau_i = A$, since τ_h can enter $\rho_i = 1$ only after $\tau_i = A$ has moved from resource $\rho_i = 1$ to $\rho_j = 2$, while arc (k, i) corresponds to giving precedence to train $\tau_h = B$. Here, w_{jh} represents the minimum time needed to release ρ_i after the entrance of τ_i in ρ_j and to make it available for τ_h. A similar discussion holds for w_{ki}.

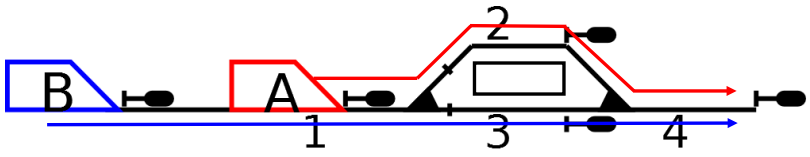

Figure 1: A simple instance with two trains.

A feasible schedule for the TD problem can be obtained by defining a complete consistent selection S and fixing all variables $t_i = l^S(0, i)$, for all $i \in N$. Then t_n is the maximum (consecutive) delay of the schedule. Therefore, the TD problem is the problem of finding a complete consistent selection S such that the length of the longest path $l^S(0, n)$ is minimum.

To summarize, the alternative graph formulation of the TD problem is a disjunctive program:

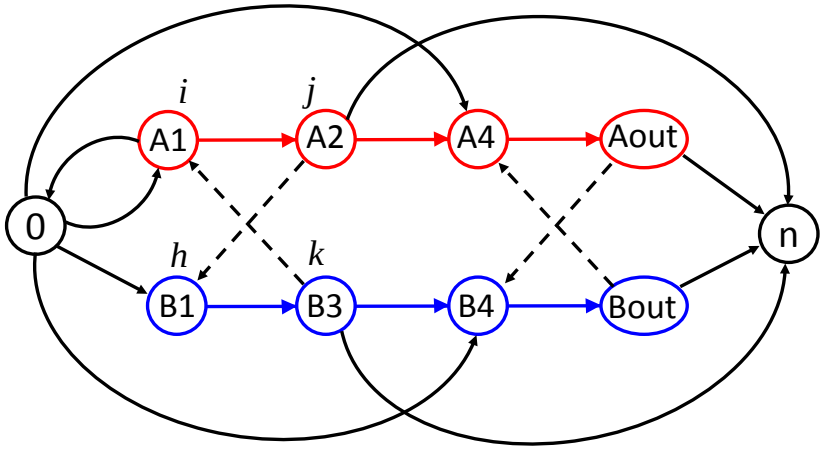

Figure 2: the alternative graph corresponding to the instance of Figure 1 (with solid fixed arcs and broken alternative arcs).

$$
\begin{array}{lll}
\min & t_n & \\
s.t. & t_j - t_i \geq w_{ij} & (i,j) \in F \\
 & (t_h - t_j \geq w_{jh}) \vee (t_i - t_k \geq w_{ki}) & [(j,h),(k,i)] \in A
\end{array}
$$

This formulation can be easily converted into a Mixed Integer Linear Program by associating a binary variable x_{jhki} to each disjunction, i.e., to each pair in A, equal to one when (j,h) is selected and equal to zero if (k,i) is selected. Then, the TD problem can be formulated as a MILP as follows:

$$
\begin{array}{lll}
\min & t_n & \\
s.t. & t_j - t_i \geq w_{ij} & (i,j) \in F \\
 & t_h - t_j + M(1 - x_{jhki}) \geq w_{jh} & [(j,h),(k,i)] \in A \\
 & t_i - t_k + M x_{jhki} \geq w_{ki} & [(j,h),(k,i)] \in A \\
 & x \in \{0,1\}^{|A|} &
\end{array}
$$

The MILP formulation can be extended to the more general TD problem in which both a route and a schedule must be defined for each train, i.e., the TD problem with routing flexibility. This can be obtained by enlarging sets F and A to contain all possible arcs for all possible train routes, and by adding for each alternative route variables $y_{uv} \in \{0,1\}$ equal to 1 if route u is chosen for train v, and 0 otherwise. Let F_u be the set of fixed arcs associated to route u. Alternative pairs are associated to all resources shared by two routes u and $\bar{u}$. Let n_T be the number of trains to be scheduled, R_v be the set of routes of train v, $R = \bigcup_{v=1,\ldots,n_T} R_v$, $A_{u\bar{u}}$ be the set of pairs of alternative

arcs associated to the resources shared by routes u and $\bar{u}$, $A = \bigcup_{u,\bar{u}\in R} A_{u\bar{u}}$. Then , the train scheduling and routing formulation is the following:

$$
\begin{array}{ll}
\min \quad t_n & \\
s.t. & \\
t_j - t_i + M(1 - y_{uv}) \geq w_{ij} & (i,j) \in F_u; u \in R_v; v = 1, ..., n_T \\
t_h - t_j + M(1 - x_{jhki}) + M(1 - y_{u\tau_h}) + M(1 - y_{\bar{u}\tau_j}) \geq w_{jh} & [(j,h),(k,i)] \in A_{u\bar{u}}; u, \bar{u} \in R \\
t_i - t_k + Mx_{jhki} + M(1 - y_{u\tau_i}) + M(1 - y_{\bar{u}\tau_k}) \geq w_{ki} & [(j,h),(k,i)] \in A_{uv}; u, \bar{u} \in R \\
\sum_{u\in R_v} y_{uv} = 1 & v = 1, ..., n_T \\
x \in \{0,1\}^{|A|}, y \in \{0,1\}^{|R|} &
\end{array}
$$

3 Algorithmic aspects

The big-M formulation introduced in the previous section provides the basis for a number of solution approaches. In principle, one could simply adopt a commercial solver to attack and solve practical instances of the big-M formulation. Unfortunately, such natural but naive approach is likely to fail on most instances of some practical interest. In fact, it is known that big-M formulations are rather weak and the typical instances of train dispatching are simply too large to be attacked directly. The solver would normally return bad quality solutions or no solution at all. To get around this difficulty different authors followed different strategies, such as embedding the big-M formulation into some smart algorithmic schema, or simply avoiding the use of big-M formulations. We quickly go through the most common options:

1. Heuristics of various type
2. Branch&Bound, with bound computed with some combinatorial methods
3. Alternative formulations
4. Decomposition methods
5. A blend of the above approaches

Concerning point 1., the literature is very vast and varied, so we refer the reader to the mentioned survey papers. As for point 2., the literature reports very few attempts, as e.g., [24] for mass transit and [10] for main line.

Proceeding with point 3., the classical alternative to the big-M formulation for scheduling problems is the so called *time-indexed* formulation, where the time horizon is discretized, and we have one binary variable y_{it} for each atomic movement (i.e. the occupation of a rail resource by a train) and each period. y_{it} is 1 if and only if the atomic movement i starts in period t. A major drawback of this approach is that it tends to introduce a huge number of binary variables and packing constraints, even for small dispatching instances. Because of the tight computing times enforced by the

application - a solution must be returned in a few seconds in order to be assessed by dispatchers in real-time - time-indexed formulations are in general preferred for off-line problems such as train timetabling. Only a few attempts have been made also in train dispatching, such as [6, 22, 27, 31]. Recently, a promising approach with only a few continuous variables and as few binary variables as for the big-M formulation has been presented in [18].

Point 4.: An alternative and quite popular technique to tackle complicated mixed integer linear programs is decomposition. The term *decomposition* basically denotes the act to replace the original problem with a sequence of smaller subproblems. The solutions to the smaller problems is then re-combined or extended to the original large problem. If one fails in this phase, some kind of *regret* mechanism must be put in place and the problem solved from start. The two most common decompositions may be seen as operating in time and space, respectively.

Decomposing in time: rolling horizon. Rolling horizon is a classical decomposition approach (see, e.g, [35, 28, 27, 37]) in which the time horizon is decomposed into smaller intervals and a subproblem is associated with each interval. Then the subproblems are solved in chronological order. At each iteration, the (part of the) solution associated with the previous subproblems is fixed, and only 'few' additional variables and constraints are left.

Decomposing in space: the macroscopic/microscopic decomposition principle. A different but still very popular type of decomposition approach is the so-called macroscopic-microscopic decomposition approach [32]. A typical two-stage implementation of this approach goes as follows: in a first stage stations are considered as points with infinite capacity, and trains are scheduled along the line(s). In the second stage, one checks if the schedule so found can be extended to an overall solution, namely reconsidering in detail the actual topology of the stations. If one fails in this phase, some of the decisions taken in the first stage must be reconsidered. Different approaches differ in the way this mechanism is implemented. Most papers resort to heuristic regret approaches [11, 12] or no approaches at all (e.g. [36]). In contrast, an exact regret mechanism is developed in in [17] and [19], based on integer programming and logic Benders' decomposition. The basic ideas goes as follows. If we let t be the scheduling vector, and x be the binary vector associated with the disjunctive terms, then the generic big-M formulation introduced in the previous section may be represented as:

$$
\begin{array}{lllll}
\min c^T t & & & & \\
s.t. & & & & \\
(i) & A_L x_L + & B_L t & \geq w_L, & \\
(ii) & & B_S t + A_S x_S & \geq w_S & \\
(iii) & & t \text{ real}, \quad x_L, x_S \text{ binary} & &
\end{array}
\tag{1}
$$

Vector x has been written as $x = (x_L, x_S)$ to distinguish between decisions associated with stations (x_S) and decisions associated with tracks between stations (x_L). Observe

that matrices A_L, A_S and vectors w_L, w_S contain in general big-M coefficients.

Next, we may identify blocks (i) and (ii) in Program (1). The two blocks only share a small subset of t variables, namely those corresponding to arrivals and departures from stations. Indeed, leaving a station amounts to entering the line between to successive stations, whereas leaving a track between two stations amounts to entering the second station. It follows that block (i) may be seen as corresponding to the macroscopic problem of controlling trains between stations, whereas block (ii) corresponds to the microscopic problem associated with stations.

In Benders' decomposition algorithm, we first solve the restricted problem (*master*) obtained by dropping block $(1.ii)$. This corresponds to taking dispatching decisions for trains running on the line, neglecting what can happen in stations. Let (x_L^*, t^*) be the optimal solution to the master problem. In order to established if t^* can be extended to a a feasible solution to $(1.ii)$ we need to solve the so called (*slave* problem). It should be apparent by now that the slave problem corresponds to solving a microscopic feasibility problem (for all stations), where arrival and departure times for all trains are fixed (by the master). If the slave problem has a solution (t^*, x_S^*) then we are done as (x_L^*, t^*, x_S^*) is an optimal solution to (1). Otherwise (x_L^*, t^*) cannot be extended to a feasible solution for the whole problem, and we identify an inequality $q^T x_L + r^T t \leq k$ which is satisfied by all of the feasible solutions to (1) but is violated by (x_L^*, t^*). Such inequality is added to the master problem and the process is iterated. A nice feature of the approach is that the slave problem further decomposes into a number of independent sub-problems, one for each station.

A different approach to space decomposition is presented in [8], where entire administrative regions (*areas* are actually collapsed into single nodes in order to carry out inter-area coordination among trains.

Finally, another type of decomposition approach consists in partitioning trains into groups of one or more elements and then solve the problem associated with each group in sequence, as in [1].

Interestingly, the three decomposition approaches introduced above mimic, in some sense, the actual behavior of human dispatchers, which somehow apply a combination of these methods. Indeed, potential conflicts are typically solved by dispatchers by neglecting the microscopic details (of the stations or of the tracks between stations). Then train movements within stations are then controlled in the necessary detail. Also, a dispatcher typically focuses on one or two trains at a time, and only for the next few movements or conflict.

4 State of practice

Even if the optimization literature has been drilling into train dispatching for over 30 years, to our knowledge there are very few operative TMSs which rely on optimization algorithms to take or suggest dispatching decisions. This gap is also highlighted in recent surveys ([3, 9] and papers such [37]. There are several reasons for this discrep-

ancy. There is little doubt that the first attempts to automatize the dispatching process resulted in disappointing failures. This was possibly due to the use of inadequate techniques (e.g., rule based systems rather than optimization models and algorithms). The unavoidable consequence was an immediate and long-lasting skepticism of railway infrastructure managers and train operators in the sheer possibility of such automation. Another factor of resistance is related to the particular relation between the two major actors of the TMS market, namely infrastructure managers and (large) system vendors. None of those actually is willing to take the risk of innovation. By one hand, developing optimization based TMSs requires large investments by the vendors, with uncertain outcomes. On the other hand, railway operators, until very recently, did not press the vendors with specific technological requirements, probably because of lack of strong motivations (such as a real competitive market) or incentives. Indeed, in Europe infrastructure managers are typically state-hold companies and the operate in monopolistic markets.

The situation seems now to be on the verge of a rapid change. A growing awareness towards the potential of optimization-based TMSs is tangible. Infrastructure managers and operators around Europe are starting to explicitly request the use of optimization modules within TMSs, as we have observed in recent tenders, e.g. Denmark, Sweden and Norway. This fact actually forces the large vendors to pursue a stronger collaboration with research centers, since developing optimization tools for train dispatching is not an easy task.

Indeed, a few, large system vendors, finally started to claim their TMSs embed some sort of optimization algorithms. This is the case, for instance, of the General Electrics Movement Planner (see GE Movement Planner 2016). Unfortunately, the only available material is a few brochures, where the company states to implement *business objective based optimization*. Similar claims are also reported in a presentation by Siemens (Eickholt 2012), but again without any significant documentation of methods and practice. Same applies to Hitachi (Hitachi 2016). Other large companies made similar claims with poor supporting documentation. Also, after some personal (but far from exhaustive) investigation, we could not identify any of the OR specialists involved in the development of such softwares, let alone scientific publications. A remarkable exception is represented by Alstom, a large multinational company headquartered in France, specialized in signaling systems and TMSs. A few years ago, in order to prepare for a large tender in Sweden where the use of optimization was explicitly required, Alstom decided to engage three academic groups with strong and long lasting expertise in railway optimization. After a sort of pre-qualification, one of these groups was selected to develop the optimization part of final product and is now collaborating with Alstom. The optimization engine is now integrated in the TMS offered by Alstom, but not yet in operation.

Indeed, according to what reported in the literature and what personally overheard or learned at specialized workshops, meetings, etc., the only few TMSs in operation actively using optimization described in the scientific literature are described in the

remaining of this section (see also [2] for more details).

In all existing systems, the optimization routines are typically embedded in a loop. First, the TMS acquires real-time information regarding the status of the network (e.g. train positions, speeds, resource availability etc). This information is fed into the optimization modules, which, combining it with the required "static" information (e.g. network layout, train connections), return one or more solutions to the current dispatching problem. Dispatchers may also "manually" interact with the systems providing further information (e.g. train delays or cancellations, network disruptions, fixed meeting points). The total time allowed for this loop is rather tight, typically a few seconds, setting a limit on the size and type of instances which can be tackled.

To our knowledge, the first "optimization-based" support system reported in the literature was embedded in a TMS developed by Bombardier Transportation, and operated in some terminal stations of the Milano Underground between 2007 and 2008 ([24]). The system was dismissed after a year because Bombardier lost a tender for a complete renewal of the TMSs controlling the lines in Milano underground system. The dispatching algorithm was an exact branch&bound. By contract clause, the system had to prove to perform better than the dispatchers in order to be actually purchased by Azienda Trasporti Milanese, the state-hold company managing Milano underground system. To this end, an intensive one-week on-field test campaign was set up in a terminal stations, which directly compared dispatchers against algorithm. Each day, four "traffic equivalent" one hour slots were identified: in two of them the algorithm had full control over the trains, while in the other two the control was assigned to the dispatchers. In spite of the small size of the station, the system performed, on average, 8% better than dispatchers both in terms of deviations from timetable and in regularity. Besides paper [24] a popular science description of this experience is presented in [25].

Again Bombardier Transportation is behind the first main line optimization based dispatching system. This was put in operation in Italy in 2011, on a singletrack regional line (Trento-Bassano del Grappa), see [23]. The use of the tool was later extended to other lines in Northern, Southern and Central Italy, such as Milano-Mortara, Piraineto-Trapani-Alcamo, Orte-Terontola-Falconara and others. The following table gives some hints about the size and geometry of such regional lines.

Line	**Stops**	**Stations**	**Length (m)**	**Tracks**
Trento - Bassano	22	14	95711	Single
Piraineto - Trapani	12	12	93532	Single
Alcamo - Trapani	14	13	116119	Single
Orte - Terontola - Falconara	54	34	283839	Mixed

Table 1: Infrastructure details. Single stands for Single-Track, Mixed stands for Single and Double-Track stands for Double-Track

Because of very strict business rules adopted by the Italian Infrastructure Manager

(RFI), the optimization algorithm embedded in Bombardiers' TMS is a heuristic. For the same reason, the dispatching loop is semi-automated: the algorithm finds alternative solutions each time it is called and presents them to the dispatcher(s) ranked by cost. Statistics show that in 94% of the cases the first solution proposed by the algorithm is accepted. It is worth noticing that such limiting business rules were introduced to help dispatchers to quickly take difficult decisions. If such rules would be ignored or dropped, then the full power of an exact optimization algorithm could be exploited. Indeed, we have tried to quantify the possible advantages to do this in [19]: it turns out that for the mentioned Orte-Terontola-Falconara line a significant average increase of trains on time (+6.2%) and reduction of trains heavily delayed (8.9%). The results (presented in Table 2) a referred to a specific week day in year 2013.

Method	**On Time**	**delay$\leq$ 10**	**delay$\leq$15**	**delay>15**
Heuristic	84.7%	86.5%	87.9%	12.1%
Exact method	90.9%	95%	96.8%	3.2%

Table 2: Exact VS Heuristic approach: comparisons on train punctuality

A new release of the above TMSs has been recently deployed in a few lines in Latvia, namely Daugavpils-Eglaine, Daugavpils-Krustpils, Rezekne-Krustpils, Zilupe-Krustpils, Karzava-Rezekne, for a total of 52 stations, with 10 communication points and 8 station gates. These lines are mainly used for freight transportation and run around 100 trains every 20 h. Again due to local business rules, the optimization algorithm is heuristic. A major innovation is that, in a particular modality, freight trains decision can be directly applied by the system without prior acceptance by dispatchers.

In a recent project involving with the Norwegian infrastructure manager (JBV) and train operating companies (NSB, FlyToget, CargoNet), an automatic dispatching system was put in operation at Stavanger control center in Norway, in February 2014. The system controlled trains on the Stavanger-Moi Line, which is 123 km long, with 16 stations, 7 line stops and 28 block sections. On weekdays, the average number of trains every 12 h is around 100. The system implements an exact, MILP algorithm described in [17, 19]. Like in previous cases, the system presents solutions in real-time to dispatchers which decide whether to accept the solutions. The system was well received by dispatchers and management. The use of the system in Stavanger for real-time dispatching was put on hold for legal reasons towards the end of 2014, because of a competitive tender issued by JBV for the renewal of the entire Norwegian signaling (and centralized traffic control) system.

Finally, concerning large stations rather than lines, the TMSs in Roma Tiburtina station (12 line points, 30 stopping points and 62 interlocking routes) and the multi-station of Monfalcone in Italy have been equipped with optimization algorithms which re-schedule and re-route trains. The optimization modules include both heuristic and

exact algorithms. Tiburtina station is the second largest station in Rome, and it is considered one of the most important and complex stations in Italy.

References

[1] Andrea Bettinellia, A Santini, and D Vigo. A real-time conflict solution algorithm for the train rescheduling problem. *under revision*, 2017.

[2] Ralf Borndörfer, Torsten Klug, Leonardo Lamorgese, Carlo Mannino, Markus Reuther, and Thomas Schlechte. Recent success stories on integrated optimization of railway systems. *Transportation Research Part C: Emerging Technologies*, 74:196–211, 2017.

[3] V. Cacchiani, D. Huisman, M. Kidd, L. Kroon, P. Toth, L. Veelenturf, and J. Wagenaar. An overview of recovery models and algorithms for real-time railway rescheduling. *Transportation Research Part B*, 63:15–37, 2014.

[4] Valentina Cacchiani and Paolo Toth. Nominal and robust train timetabling problems. *European Journal of Operational Research*, 219(3):727 – 737, 2012. Feature Clusters.

[5] Gabrio Caimi, Fabián A. Chudak, Martin Fuchsberger, Marco Laumanns, and Rico Zenklusen. A new resource-constrained multicommodity flow model for conflict-free train routing and scheduling. *Transportation Science*, 45(2):212–227, 2011.

[6] Gabrio Caimi, Martin Fuchsberger, Marco Laumanns, and Marco Lüthi. A model predictive control approach for discrete-time rescheduling in complex central railway station areas. *Computers & Operations Research*, 39(11):2578–2593, 2012.

[7] Francesco Corman, Andrea D Ariano, Alessio D. Marra, Dario Pacciarelli, and Marcella Samà. Integrating train scheduling and delay management in real-time railway traffic control. *Transportation Research Part E: Logistics and Transportation Review*, 2016. In press.

[8] Francesco Corman, A D'Ariano, D Pacciarelli, and M Pranzo. Optimal inter-area coordination of train rescheduling decisions. *Transportation Research Part E: Logistics and Transportation Review*, 48(1):71–88, 2012.

[9] Francesco Corman and Lingyun Meng. A review of online dynamic models and algorithms for railway traffic management. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1274–1284, 2015.

[10] Andrea D'Ariano, Dario Pacciarelli, and Marco Pranzo. A branch and bound algorithm for scheduling trains in a railway network. *European Journal of Operational Research*, 183(2):643 – 657, 2007.

[11] Twan Dollevoet, Francesco Corman, Andrea D'Ariano, and Dennis Huisman. An iterative optimization framework for delay management and train scheduling. *Flexible Services and Manufacturing Journal*, 26(4):490–515, 2014.

[12] Twan Dollevoet, Dennis Huisman, Leo Kroon, Marie Schmidt, and Anita Schöbel. Delay management including capacities of stations. *Transportation Science*, 49(2):185–203, 2014.

[13] W. Fang, S. Yang, and X. Yao. A survey on problem models and solution approaches to rescheduling in railway networks. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):2997–3016, Dec 2015.

[14] N Ghaemi and RMP Goverde. Review of railway disruption management practice and literature. In *6th International Conference on Railway Operations Modelling and Analysis-RailTokyo2015,*, 2015.

[15] IA Hansen. *State-of-the-art of railway operations research*, chapter A 4, pages 35–47. Timetable Planning and Information Quality. WIT Press, 2010.

[16] Steven Harrod and Thomas Schlechte. A direct comparison of physical block occupancy versus timed block occupancy in train timetabling formulations. *Transportation Research Part E: Logistics and Transportation Review*, 54:50 – 66, 2013.

[17] L. Lamorgese and C. Mannino. An exact decomposition approach for the real-time train dispatching problem. *Operations Research*, 63(1):48–64, 2015.

[18] Leonardo Lamorgese and Carlo Mannino. A non-compact formulation for job-shop scheduling problems in transportation. In Leo G. Kroon, Anita Schöbel, and Dorothea Wagner, editors, *Algorithmic Methods for Optimization in Public Transport (Dagstuhl Seminar 16171) Dagstuhl Reports*, volume 6, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[19] Leonardo Lamorgese, Carlo Mannino, and Mauro Piacentini. Optimal train dispatching by benders'-like reformulation. *Transportation Science*, 50(3):910–925, 2016.

[20] Richard Lusby, Jesper Larsen, Matthias Ehrgott, and David Ryan. Railway track allocation: models and methods. *OR Spectrum*, December 2009.

[21] Richard Lusby, Jesper Larsen, David Ryan, and Matthias Ehrgott. Routing trains through railway junctions: A new set-packing approach. *Transportation Science*, 45(2):228–245, 2011.

[22] Richard M Lusby, Jesper Larsen, Matthias Ehrgott, and David M Ryan. A set packing inspired method for real-time junction train routing. *Computers & Operations Research*, 40(3):713–724, 2013.

[23] C. Mannino. Real-time traffic control in railway systems. In Alberto Caprara and Spyros Kontogiannis, editors, *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, volume 20 of *OpenAccess Series in Informatics (OASIcs)*, pages 1–14, Dagstuhl, Germany, 2011. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[24] C. Mannino and A. Mascis. Optimal real-time traffic control in metro stations. *Operations Research*, 57:1026–1039, 2009.

[25] Carlo Mannino and Alessandro Mascis. Fast track to fixing rail delays-award-winning automated rail re-routing system saves time and money. *OR MS Today*, 37(2):28, 2010.

[26] A. Mascis and D. Pacciarelli. Job shop scheduling with blocking and no-wait constraints. *European Journal of Operational Research*, (143 (3)):498–517, 2002.

[27] Lingyun Meng and Xuesong Zhou. Simultaneous train rerouting and rescheduling on an n-track network: A model reformulation with network-based cumulative flow variables. *Transportation Research Part B: Methodological*, 67:208–234, 2014.

[28] Lars Kjær Nielsen, Leo Kroon, and Gábor Maróti. A rolling horizon approach for disruption management of railway rolling stock. *European Journal of Operational Research*, 220(2):496–509, 2012.

[29] P. Pellegrini, J. Marlière, and D. Rodriguez. Optimal train routing and scheduling for managing traffic perturbations in complex junctions. *Transportation Research Part B*, 59:58–80, 2014.

[30] B. Roy and B. Sussmann. Les problèmes d'ordonnancement avec contraintes disjonctives, note d.s. no. 9 bis. Technical report, SEMA, France, 1964.

[31] Güvenç Şahin, Ravindra K Ahuja, and Claudio B Cunha. Integer programming based solution approaches for the train dispatching problem. 2010.

[32] Thomas Schlechte, Ralf Borndörfer, Berkan Erol, Thomas Graffagnino, and Elmar Swarat. Micro–macro transformation of railway networks. *Journal of Rail Transport Planning & Management*, 1(1):38–48, 2011.

[33] Ambra Toletti and Ulrich Weidmann. Modelling customer inconvenience in train rescheduling. In *16th Swiss Transport Research Conference (STRC 2016)*. Swiss Transport Research Conference (STRC), May 2016.

[34] Johanna Törnquist. Computer-based decision support for railway traffic scheduling and dispatching: A review of models and algorithms. In Leo G. Kroon and Rolf H. Möhring, editors, *5th Workshop on Algorithmic Methods and Models for*

Optimization of Railways (ATMOS'05), volume 2 of *OpenAccess Series in Informatics (OASIcs)*, Dagstuhl, Germany, 2006. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[35] Johanna Törnquist. Railway traffic disturbance management:an experimental analysis of disturbance complexity, management objectives and limitations in planning horizon. *Transportation Research Part A: Policy and Practice*, 41(3):249 – 266, 2007.

[36] Shuguang Zhan, Leo G Kroon, Lucas P Veelenturf, and Joris C Wagenaar. Real-time high-speed train rescheduling in case of a complete blockage. *Transportation Research Part B: Methodological*, 78:182–201, 2015.

[37] Shuguang Zhan, Leo G Kroon, Jun Zhao, and Qiyuan Peng. A rolling horizon approach to the high speed train rescheduling problem in case of a partial segment blockage. *Transportation Research Part E: Logistics and Transportation Review*, 95:32–61, 2016.