Single European Sky and Single European Railway Area: A System Level Analysis of Air and Rail Transportation
Paola Pellegrini, Joaquin Rodriguez

To cite this version:

HAL Id: hal-00870632
https://hal.archives-ouvertes.fr/hal-00870632
Submitted on 7 Oct 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

Air and rail transportation systems are characterized by important common features: they serve a significant share of passenger traffic in Europe; their functioning relies on the cooperation of many stakeholders operating a fixed timetable, often with competing objectives; and they have been characterized by quite a fragmented development following national borders. For these systems, the European Commission envisages a common future in terms of an increase of efficiency and elimination of national borders. In this paper, we analyze the two systems in the perspective of their common development and we underline the main existing differences linked to the management of the infrastructure. These differences concern the processes implemented for specifying and allocating capacity, both in strategic planning and in real-time intervention. Our analysis suggests that the air transportation system is moving in the direction indicated by the European Commission, as well as the rail transportation system for what concerns international train paths (typically high-speed trains). However, a substantial separation still exists between conventional rail transportation systems of different countries.
Keywords: air transportation; rail transportation; Single European Sky; Single European Railway Area; system capacity management.

1 Introduction

Air and rail are two extremely important, and sometimes competing, means of transportation. Air transportation allows very long distances to be covered in a rather short time. Rail transportation allows rather long distances to be covered in a reasonable amount of time at a relatively low economic and environmental cost. According to Eurostat (2009), in 2006 air traffic served 8.6% of passengers per kilometer in Europe, and rail traffic 6.1%. The corresponding figures for freight are 0.1% and 10.5%, measured in tonnes per kilometer. Although road traffic serves a much higher percentage of both passenger and freight, its great environmental impact necessitates the investigation for effective alternatives.

Due to historical and institutional reasons, European air and rail transportation systems are managed, to some extent, at a national level, with sometimes significant differences between countries. The consequent fragmentation imposes evident limitations on the development of the systems.

For coping with the need of efficient European transportation systems, the European Commission has launched two ambitious programs: the Single European Sky (European Commission, 1999) and the Single European Railway Area (European Commission, 2010a). The Single European Sky establishes targets in key areas of safety, network capacity, effectiveness and environmental impact for what concerns the European air transportation system. Moreover, it envisages a technological modernization and optimization of the system through the Single European Sky ATM Research (SESAR) Programme (SESAR Consortium, 2008), where ATM stands for Air Traffic Management. The Single European Railway Area intends to create a unique efficient inter-European freight and passenger market, favoring cross-border competition.

In the optic of this common target, we propose an analysis of the main peculiarities of air and rail transportation systems. In particular, we consider the main elements and processes in the two systems for what concerns the infrastructure management, and we report the main legislation and research trends which concern them. This analysis allows the identification of similarities and differences between the two systems, and of the level of development of the European homogenization process. Other studies focus on the two transportation systems together, but from very different perspectives. When this happens, the main focus is typically the identification and the tentative explanation of market shares of air and high-speed rail transportation (Adler et al., 2010; Buckeye, 1992; Dobruszkes, 2011; Ehrenberger and Fischer, 2011; Janich, 2003; López-Pita and Robusté Anton, 2003; Park and Ha, 2006; Román et al., 2007; Steer Davies Gleave, 2006). Other studies (Chen and Cheng, 2010; Davidsson et al., 2005) propose a review of agent based techniques applied to different transportation modes, including air and rail transportation.

The rest of the paper is organized as follows. Section 2 discusses the main
steps which have characterized the evolution of air and rail transportation systems from a historical and institutional point of view. Section 3 describes the main elements composing the infrastructure of the two systems. Section 4 presents the process implemented for allocating infrastructure capacity in the strategic planning, which typically takes place months or years in advance (Barnhart et al., 2012). Section 5 details the research trends and current practice for what concerns real-time intervention. Finally, Section 6 concludes the paper.

2 Historical and institutional framework

Historically, both air and rail transportation systems have developed in a rather fragmented fashion, mostly following national borders. The development of the two systems has been negatively impacted in distinct ways as a result of this fragmentation. Moreover, different institutional roles have been attributed to the two systems: air transportation is mostly a commercial framework, and rail transportation is often intended as a service to be made available to the population. In this section, we describe the evolution of the systems and the main developments which are currently envisaged.

2.1 Air transportation

In Europe, commercial aviation started in the 1920’s, with private airlines benefiting from governmental subsidies. The first commercial flights were performed between London and the Continent by four airlines: a Dutch, two French and a Belgian airline. Over the years, the market has grown markedly: annually, commercial airlines operate approximately 9.5 million flights in Europe (Eurocontrol Performance Review Commission, 2010). Besides commercial airlines, air traffic is populated by military and general aviation operators, such as private individuals operating their own aircraft.

The definition of an air traffic control system came with the increase of the number of aircraft populating the sky. The International Convention for Air Navigation (also known as the Paris Convention), established in 1919, stated the sovereignty of each state over its own airspace. Moreover, it declared that “Every aircraft in a cloud, fog, mist, or other condition of bad visibility shall proceed with caution, having careful regard to the existing circumstances”. At that time, the only aid to pilots was represented by flags waved at airports to signal permission for landing or take-off, and successively by signaling lamps located at strategic positions. In Europe, the first modern air traffic control center was the control tower at Schiphol airport, in the Netherlands, thanks to the modernization brought about in occasion of the Olympic Games held in Amsterdam in 1928. Here, after its complete destruction which occurred during the Second World War, the control tower was equipped with a control radar, and controllers employed novel techniques for sequencing landing and take-offs (Cook, 2007). Since then, general air traffic control (ATC) rules quickly spread throughout Europe, even if the equipment of the control centers and some
specific criteria employed are still managed at a national level (see Section 4.1).

Nowadays, five main direct stakeholder groups operate within the air transportation system (SESAR Consortium, 2008):

1. the end-user customers: passengers and freight;
2. the airspace users: commercial airlines, military, general aviation (all flights other than military and scheduled airline and regular cargo flights) and business aviation (any general aviation aircraft used for a business purpose);
3. the air navigation service providers (ANSPs): providers of air traffic control, airspace management and air traffic flow management services. In Europe, these services are mostly provided by national ANSPs, typically one per State. The provision of some services has been delegated to a single European organization, namely EUROCONTROL;
4. the aerodrome community: airports and aerodromes, including major hub airports, regional airports, local aerodromes and military airfields;
5. the supply industry: aircraft manufacturers, suppliers of supporting systems.

In this paper, we will focus on the stakeholders which have a major impact on the air transportation system as a whole and whose behavior may be to some extent controlled through regulation: airlines, ANSPs and airports. In particular, we will focus on how they cooperate for efficiently exploiting the system capacity.

ANSPs are often managed by national civil aviation authorities, although they should be managed by independent entities (International Civil Aviation Organization, 1996). European airports are mostly private (Cook, 2007). For airlines, the situation is more controversial. They are private companies, and governmental subsidies have been officially forbidden since 1957, by the Treaty of Rome. However, the treaty forbade subsidies which distorted or threatened to distort competition. This statement has not been sufficient to eliminate all governmental interference within the air transportation market. However, airlines in Europe are mostly private, possibly partly owed by governments; operational decisions such as pricing and route selection are, in principle, managed according to commercial evaluations rather than political reasons. In the past, the lack of application of the commercial logic in favor of a more politically oriented one has lead to airline bankruptcies, as discussed by Beria et al. (2011) regarding the case of Alitalia, in Italy.

According to the European Commission (2008b), Member States’ aviation authorities are in charge of assessing air carriers’ financial stability for obtaining and preserving an operating license. The assessment performed, nonetheless, does not appear to be uniform across states (Simonelli and Caroli, 2013).

The coexistence of private stakeholders, each operating maximizing its own profit, implies a high level of fragmentation in the current system. This fragmentation is evident, for example, in the way airport capacity is allocated in
the strategic planning described in Section 5.1, and most of all in the lack of transparency behind stakeholder decisions.

2.2 Rail transportation

The birth of modern rail transportation in Europe can be dated back to the 1820’s, when the first steam locomotives appeared in England (Hylton, 2007). The first public passenger railway was built between two cities named Witon Park and Stockton-on-Tees via Darlington, and is commonly known as the Stockton and Darlington railway. The first regulation concerning railways dates back to 1840, when the English government approved the Regulation of Railways Acts imposing that no railway could be opened without notice, that railway companies should make returns, and that the Board of Trade could authorize inspections on the railways. Originally, railways in England were owned by four private companies: the Great Western Railway, the London, Midland and Scottish Railway, the London and North Eastern Railway and the Southern Railway. The situation changed in 1947, at a time when rail transportation had already been recognized as a public service: railways were nationalized into the British Railway through the Transport Act.

In other European countries, railways followed similar paths as in England, starting with regional networks owned by private companies and ending with nationalization: in France, Société Nationale des Chemins de fer Français (SNCF) was created in 1938; in Germany, Deutsche Reichsbahn was created in 1920; in Spain, RENFE and FErrocarriles de Vía Estrecha (FEVE) were created in 1941 and 1965, respectively; in Italy, Ferrovie dello Stato was created in 1905. These national operators held the monopoly of the market in their respective countries.

From the 1990’s, the European rail transportation system started reverting to privatization to increase the efficiency of the service delivered. The European Commission promulgated in 1991 a directive imposing countries to separate the management of the railway infrastructure and of the operative service. This separation is commonly referred to as unbundling and has been undertaken in different ways and at different times in the various European countries. Typically, both the infrastructure managers and the railway undertakings are under some level of control of transportation authorities; a remarkable exception is represented by railway undertakings in England, which operate through a franchising system (Nash, 2008): in this way, the transportation authority achieves a high level of control over the service and over its costs, while enhancing competition.

Unbundling is not necessarily the best way to enhance competition (Gómez-Ibáñez and De Rus, 2006), mostly because its actual realization is very difficult in the practice. In fact, there is a great need for coordination between infrastructure managers and railway undertakings. This may strongly favor incumbent railway undertakings by reducing transaction costs in terms of the negotiation and enforcement of contracts between the train operator and the infrastructure manager (Merkert, 2009). In addition, contracts between railway undertakings
and infrastructure managers (see Section 2.2) are typically of too short duration to provide a suitable return on investment for new entrant railway undertakings (Nash, 2010). Nonetheless, no valid alternatives have been envisaged so far.

Despite the current movement toward a competitive market, the rail transportation system still includes few large railway undertakings operating mostly in one country. This is particularly true for passenger conventional rail that addresses its service to a very fragmented market, where the railway undertakings offer a service defined \textit{a priori}, and many individuals decide whether to consume it or not. A different trend is developing for high-speed lines and freight transport. The difference in the market structure is due to the fact that high-speed lines constitute a rather new market, which often exploits a dedicated infrastructure. It is also not unusual to have high-speed trains crossing national borders: the novel European vision has quite an easy access to regulation in this framework. As for what concerns the market of freight trains, it operates on a service-on-demand basis: typically, freight transportation is organized only when there is a demand. Hence, the passenger and freight markets develop at different times, typically with passenger transport organization preceding freight transport one. In this sense, there is not a real competition for the use of the infrastructure among railway undertakings operating passenger transport and railway undertakings operating freight transport. In this context, the few large railway undertakings present in each European country operate in a sort of virtual monopoly.

Overall, the main stakeholders which can be identified in the rail transportation system are:

1. the end-user customers: passengers and freight;
2. the railway undertakings;
3. the infrastructure managers;
4. the supply industry: train manufacturers, suppliers of supporting systems.

As for the air transportation system, in this paper we will focus on the stakeholders which have a major impact on the system and whose behavior may be to some extent controlled through regulation: railway undertakings and infrastructure managers. In particular, we will focus on the processes which are currently implemented, or envisaged for the future, and which concern the way in which they access and exploit the system.

3 Description of the infrastructure

The historical and institutional evolution of both air and rail transportation systems in Europe brought a significant level of fragmentation and a scarce competition, which limit the efficient growth of the systems. For tackling these issues, the European Commission is pushing on the elimination of national barriers, both material and immaterial. This is particularly important for the two
systems for different reasons: for air transportation, the elimination of national barriers is necessary to achieve a high level of efficiency for coping with the foreseen increase of demand in the near future (Eurocontrol Experimental Centre, 2008); for rail transportation, it is necessary for eliminating the virtual monopoly that characterizes national markets for conventional passenger traffic, and that clearly limits the efficiency of the system. In fact, by eliminating national borders, railway undertakings of significant sizes may compete directly with each other, rather than only with much smaller and younger companies.

The elimination of national barriers comes, first of all, from the homogenization of the infrastructures across countries. Although the infrastructure characterizing the air and the rail transportation systems are fundamentally different, it is possible to identify some theoretical similarities between them. In particular, here we focus on the elements of the infrastructure which have an impact on the specification of the capacity of the systems and on their traffic management.

3.1 Air transportation system

The infrastructure of the air transportation system includes airports as well as en-route airspace. While this definition is intuitive for airports, it may be rather unexpected for en-route airspace. However, as airports, en-route airspace has a limited capacity that must be respected for ensuring safety, and the use of which must be properly managed for enhancing the efficiency of the system. The available airspace is the area included between a minimum and a maximum flight level, with the exception of areas around airports. These bounds are often different in different countries (Cook, 2007), forcing pilots to adjust their flight plans when crossing borders. En-route airspace is split into sectors for allowing a well-defined responsibility identification in traffic management. Figure 1 shows how the space is organized along a flight route.

Airports consist of several subsystems, such as runways (where aircraft land or take-off), taxiways (that aircraft use for reaching and leaving runways), apron stands (where aircraft stop), and terminals (buildings where passengers transfer between ground transportation and airport facilities), each with its own capacity limitations. Figure 2 depicts these main airport elements.

3.1.1 Capacity specification

At major airports, the capacity of the set of runways is the most restricting element in the great majority of cases (Ball et al., 2007). The capacity of a single runway depends on many factors, the most important of which are:

1. the mix of aircraft classes using the airport. EUROCONTROL typically distinguishes three classes: heavy, medium and light aircraft;

2. the separation requirements imposed by the ATM system. They can be distinguished into wake vortex separation and airborne separation. Wake
vortex separation\(^\text{1}\) protects light aircraft from the hazards posed by the wake vortices generated by heavier aircraft. The minimum wake vortex separation distance between aircraft in case of solely arrival movements is reported in Table 1 as an example (International Civil Aviation Organization, 1996). For the non-listed pairs of leading and following aircraft, no separation is imposed due to aircraft classes. Airborne separation ensures that the runway is occupied by only one aircraft at any time;

3. the type and location of exits from the runway. They can be either conventional or high speed: a conventional exit forms a 90° angle with the runway and requires pilots to slow down the aircraft considerably (up to

\(^{1}\)While flying, wake turbulence forms behind an aircraft (Holzapfel and Gerz, 2012). This turbulence is particularly hazardous behind taking-off or landing aircraft, due to the high angle of attack. For avoiding such hazards, a minimum separation is imposed to aircraft taking-off or landing at an airport.
Table 1: Minimum wake vortex separation between aircraft of different classes.

<table>
<thead>
<tr>
<th>leading aircraft</th>
<th>following aircraft</th>
<th>separation distance (nautical miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>heavy</td>
<td>heavy</td>
<td>4</td>
</tr>
<tr>
<td>heavy</td>
<td>medium</td>
<td>5</td>
</tr>
<tr>
<td>heavy</td>
<td>light</td>
<td>6</td>
</tr>
<tr>
<td>medium</td>
<td>light</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 3: Representation of a sector, which includes airways and way-points.

about 15 km per hour); a high speed exit forms a smaller angle with the runway and permits the turn at a speed of about 100 km per hour;

4. the mix of movements on the runway (arrivals only, departures only, or mixed) and the sequencing of the movements;

5. the weather conditions, in terms of visibility, cloud ceiling, and precipitation;

6. the technological state and overall performance of the ATM system.

Given the large number of variables involved in the capacity specification, the precise airport capacity is hardly ever specified. In the literature, several mathematical models have been developed over the years for specifying the capacity of a single runway under different sets of conditions (Barnhart et al., 2003).

Capacity restrictions within the airspace are often associated with sectors. The capacity of a sector is specified as the maximum number of aircraft that can safely enter a sector in a given time horizon. Typically, controllers introduce a structure in a sector for being able to properly deal with several aircraft simultaneously. For example, aircraft may be routed along specific paths (airways) and only a few way-points may be used for entering and exiting a sector. Figure 3 shows these main sector elements.

The specification of the capacity of a sector depends on the workload that can be borne by controllers without relaxing safety requirements. Hence, no mathematical models can be produced for this specification. The literature in this field is mostly related to the analysis of the factors which affect the com-
plexity of the controller’s task, and on their impact on the controller’s workload (Hinston et al., 2002; Manning et al., 2002).

3.1.2 Capacity management

In the strategic planning, capacity is handled by airport managers. They base capacity allocation on the so-called theoretical capacity: a fixed quantity of slots is considered available for each airport when operating in normal conditions. A slot is the permission given to an airline to use the full range of airport infrastructure necessary to operate on a specific date and time for landing or taking-off (European Commission, 1993) at a congested airport. Typically, the step used for time discretization in the definition of slots is of some minutes (of the order of ten minutes (Kösters, 2007)). Many elements, such as the specific aircraft used, have an impact on actual capacity. However, theoretical capacity is computed based on standard hypothesis. For example, the theoretical capacity of Munich airport (Germany) is 89 aircraft per hour (Kösters, 2007).

Airport ordinary maintenance is normally performed off hours without impacting airport operations. Hence, no planned capacity limitations normally exist due to maintenance. When extraordinary maintenance is required, for example when runways need to be rebuilt, an airport may be closed for a few months.

When multiple runways exist at an airport, they can be either parallel or near-parallel, or crossing (International Civil Aviation Organization, 1999). Parallel and near-parallel runways can be used concurrently, respecting minimum separation constraints. The number of runways concurrently in use at an airport operating in normal conditions is a constant value. This number is used for computing theoretical capacity.

In the strategic planning, the airspace capacity is not considered at all.

During real-time intervention, aircraft movements are managed by ANSPs, both for what concerns airports and en-route space. Different tasks may characterize the management of each type of movement, and hence controller teams with different specialization exist. In particular, ground control is in charge of managing movements on apron stands and taxiways, local control has the responsibility on runways and in the air near the airport (typically 5 to 10 nautical miles, or 9 to 18 km), approach and terminal control deal with movements in the en-route airspace in the immediate vicinity of airports (typically the airspace in a radius of 30 to 50 nautical miles, or 56 to 93 km, from the airport). Local controllers also select the runways in use at the airport (International Civil Aviation Organization, 1999). The choice is based on factors as wind strength and direction, noise emissions, traffic patterns, and local regulations.

For using capacity, airlines taking-off or landing at an airport must pay a charge. The definition of the amount of this charge belongs to each single airport, which has to follow nationally imposed rules and, since 2011, further indications by the European Commission (2009b). According to the novel European regulation, airports handling more than five million passengers per year (or in any case the largest airport in each Member State) must ensure: trans-
parency on the costs which charges are to cover; non-discrimination, that is, airlines receiving the same service must pay the same charge; the existence of systems of consultation on charges between airports and airlines. Airport charges are used for covering maintenance and operational costs. The community is not unanimous on the suitability of the European regulation in this framework. For example, Starkie (2012) criticizes the imposition of a common procedure of charge definition. He claims that an actual market based on the existing relationships between airlines and airports would be more efficient in this framework.

Also the use of en-route airspace implies the payment of charges: a route charge is imposed to each flight actually performed. Until the end of 2011, ANSPs worked on a full cost recovery basis: ANSPs recovered the cost for facilities and services provided to airspace users by imposing charges taking into account the distance flown and, less than proportionately, the aircraft weight. Starting from January 2012 a regulation of the European Commission (2010b) states that risks and opportunities in the ATM network must be shared by ANSPs and airspace users. To this aim, Member States, assisted by the ANSPs involved, fix charges in a transparent manner, after consulting the airspace users representatives on determined costs, planned investments, service unit forecasts, charging policy and resulting unit rates. Some literature is devoted to the study of novel possible methods for computing route charges (Castelli et al., 2005, 2003; Raffarin, 2004; Castelli et al., 2013), and on their possible impact on air traffic.

3.2 Rail transportation

The infrastructure in the rail transportation system is represented by the railway network. In this paper we follow the terminology reported by Pachl (2008), and we focus only on the pieces of infrastructure which have an impact on the specification and allocation of capacity or on the traffic management.

The railway network is first of all made of tracks on which trains may run. At a macroscopic level, the network includes nodes and links between nodes. Links between nodes are called lines. Sets of lines connecting multiple nodes in a row are called routes. Figure 4 shows a bi-dimensional representation of a railway network seen at a macroscopic level.

In the railway network, nodes correspond to junctions in which trains can cross and change their order. According to this definition, a station is a particular junction in which passengers may get on and off trains, and freights may be loaded and unloaded. In this paper, we are not interested in any issue related to stations, but for what concerns train movements.

At a microscopic level, train movements on the infrastructure are regulated through a signaling system and a well defined set of rules. The traditional system is based on light signals separating portion of tracks called block sections. Several variants exist for the signaling system, typically depending on the country (Hotchkiss, 1995). In general terms, a signaling system has $n$ possible aspects. In the most common configuration, $n$ is equal to three: green
aspect, that allows the driver to proceed at the scheduled speed; red aspect, that imposes, according to the local regulation, either the stop or the travel at visibility speed, i.e., a speed such that the driver is able to completely stop the train if he sees another train in front of him; yellow aspect, that indicates to the driver that he shall be able to stop the train within the next block section. When \( n \) is greater than three, other restrictive aspects (e.g., flashing yellow) exist: the first restrictive aspect indicates that the driver must be able to stop the train within \( n - 2 \) block sections.

The block sections delimited by signals may contain sequences of smaller track-sections, known as track-circuits, on which the presence of a train is detected through an electrical detection system (De Vilder, 1995). Multiple block sections may share some track-circuits. The default aspect of a signal is red. To allow a train movement entering a block section, the signaling system must detect that all its track-circuits are clear. If the clear condition is not verified, a red light is signaled at the beginning of the block section itself. Figure 5 depicts a schematic representation of the infrastructure of a simple junction, considering the particular case of a station. Here, track-circuits are indicated as tc and signals are indicated as s, both indexed with a progressive number. A signal allows train movement on a block section in a precise direction: for example, signal s1 allows movements in the block section including tc1, tc2 and tc7, in this order, and signal s5 in the one including tc4 and tc3, in this order. Platform tracks, or more simply platforms, allow passengers to get on and off the trains.

A typical representation at the microscopic level of a train path is the so
called *blocking-time stairway* (Pachl, 2002) on a time/distance plot, as shown in the top plot of Figure 6. Here, space is represented on the horizontal axis in terms of distance from a given point. While moving either from or toward the reference point, the train traverses subsequent block sections. Each block section is reserved for a train for a duration equal to the *blocking-time*. It results from the sum of: the time for forming the route, the time for the driver to see the signal and react, the approach time for reaching the signal, the time between the current and the following signal, the clearing time for exiting the block section, and the release time to unlock the block system (Figure 6, bottom plot). The blocking-time for which the block section is not available for any other train is represented on the vertical axis. In the blocking-time stairway, a rectangle represents the blocking-time for each block section used by the train.

In Europe, the signaling systems are progressively being substituted by the European Rail Traffic Management System (ERTMS) (European Commission, 2012a), in which the traffic control system is mounted in trains driving cabins,
and communications are based on a GSM radio mobile system. However, the deployment of ERTMS at an European level is far from being accomplished (European Commission, 2009a).

3.2.1 Capacity specification

In the rail transportation system, the capacity of a piece of infrastructure is defined as the maximum number of trains that can be operated on it in a given time horizon (U.I.C. International Union of Railways, 2004). As for airports in the air transportation system, capacity of portions of the railway network depends on the characteristics of the movements considered, notably on their number, on their direction, on the average speed maintained, on the stability warranted in terms of time supplement, and on the heterogeneity of the trains used. In this context, times supplements represent margins added to the expected train traveling time during the timetable design. These supplements are inserted to increase the robustness of the timetable, as explained in Section 4.2.

The U.I.C. International Union of Railways (2004) describes the major constraints which must be considered when specifying capacity. Namely, it points out:

1. Priority: existing priority regulations that specify or influence the mix of trains to be performed;

2. Timetable structure: for example, integrated regular-interval timetable or line systems (U.I.C. International Union of Railways, 2004);

3. Capacity allocation process: the capacity allocation process may be different in different countries, and it may have a noticeable impact on the capacity available along lines used on international train paths;

4. Design rules: the train path duration may be influenced by the design rules in different countries, and this may have an impact on the capacity of lines used on international train paths;

5. Environmental protection: rules for environmental protection may influence both the specific train movements and the volume of traffic, for example, for reducing noise emissions;

6. Safety aspects and technical constraints: according to the characteristics of a line (for example, to the presence of tunnels) and of the trains operated (for example, trains transporting dangerous goods), these constraints may have a great impact on the capacity specification;

7. Theoretical capacity: it can be computed by considering absolute train-path homogeneity, shortest possible spacing between trains, and national regulation as only normative conditions.

The problem of specifying capacity may be tackled in terms of specification of the residual capacity with respect to an existing timetable. This specification
is made through the solution of the saturation problem, that is, the assessment of the number and characteristics of train movements that might be added to the timetable without violating any safety or operational constraint. The saturation problem can be formulated as the problem of routing and scheduling the additional trains. In the academic literature, some studies focus on the solution of this problem. Notably, Delorme et al. (2001) propose two heuristics for tackling the saturation problem. The first one is based on a constraint programming approach, the second one on a GRASP metaheuristic algorithm for tackling a unicost set packing model of this problem. The two approaches are further developed by Delorme et al. (2004) and Rodriguez et al. (2002), respectively. Burkolter (2005) specifies residual capacity through a Petri net approach. Libardo et al. (2011) propose an integer programming formulation for solving a simplified version of the saturation problem exactly. In this simplified version of the problem, train diversity is neglected. All these analyses are based on the principles stated by U.I.C. International Union of Railways (2004) on how infrastructure capacity shall be specified. Although agreed alternatives have not been proposed, these principles are not unanimously accepted, as discussed by Lindner (2011).

3.2.2 Capacity management

The infrastructure manager handles capacity both in the strategic planning and during the real-time intervention.

In the strategic planning, when railway undertakings design the timetable, they request train paths to the infrastructure manager for the identified services. As capacity is not sufficient for satisfying demand, the infrastructure manager must decide how to allocate it. Moreover, he must ensure that maintenance is possible along the network. In fact, the railway infrastructure needs a rather frequent maintenance for being safely usable at maximum speed. Typically, the infrastructure managers schedule all maintenance works in some predefined time periods, for example at night. Another possibility for scheduling maintenance works emerges in case a line includes separate track sections normally used for trains traveling in opposite directions. In this case, it may be possible to schedule maintenance on one track section and use the other one in both directions. The capacity of the only open track section is in this case much lower than the original capacity, as exemplified in Figure 7 through blocking-time stairways. The left picture represents the fully operational situation: nine trains are scheduled and superimposition of blocking-time stairways is not important, since they concern different block sections. The right picture represents the case of closure of a track section for maintenance: only five trains can be scheduled, since we must have one train leaving the shared track section before a train going in the opposite direction can enter it.

As for the infrastructure, frequent maintenance must be scheduled for the rolling-stock. Typically, such a maintenance is planned during the timetable design. Hence, it impacts the possible timetable rather than the available capacity. In the literature, some methods have been proposed to deal with the rolling-
stock maintenance jointly with the timetable design (Liebchen and Peeters, 2002; Cadarso and Marín, 2012).

During the real-time intervention, dispatchers perform the specific tasks of imposing train ordering and routing in the network for minimizing delays.

Similarly to the air transportation case, railway undertakings must pay a charge for using the infrastructure. The computation of this charge may change from country to country. A directive of the European Commission (2001) and its subsequent amendments (European Commission, 2002a, 2004a, 2007b) fix some general rules which should be respected by all European countries, with the exception of Ireland, Northern Ireland, Greece and Luxembourg. According to this regulation, charges must be payed to the infrastructure managers and they must be used for funding their business. In principle, the charge imposed to each train should be equal to the cost directly borne by the infrastructure manager for ensuring the good condition of the pieces of infrastructure used by the train itself. In addition, if capacity on a line is scarce, a higher charge may be imposed for using it. Finally, charges may include a further amount due to the environmental impact of operating the trains. According to the regulation, moreover, “the charging and capacity allocation schemes should permit equal and non-discriminatory access for all undertakings, and should attempt, as far as possible, to meet the needs of all users and traffic types in a fair and non-discriminatory manner” (European Commission, 2001).

3.3 Main similarities and differences

The infrastructure of the air and rail transportation systems have some structural similarities. In particular, in air transportation, aircraft stop at airports and travel along routes through sectors, and in rail transportation, trains stop at stations and travel along routes composed of lines connecting junctions. Capacity specification and allocation at airports and stations are particularly difficult and critical for the performance of the systems. In both systems, both capacity specification and allocation shall be referred to a specific schedule, due, for example, to weight, route and speed considerations.
For making a parallelism of the infrastructure characterizing the air and the rail transportation systems, at a macroscopic level we may interpret lines in the rail system as en-route sectors and nodes as terminal airspace: here a flight can start or terminate its journey by leaving or reaching the airport, or it can continue traveling; its route must in any case be followed with great attention by air traffic controllers. At a microscopic level, we may see stations as airports. Here, platforms correspond to apron stands and critical switches where trains may pass for entering and exiting the station correspond to runways: a certain separation must be ensured between consecutive aircraft using a runway, and a certain separation must be ensured between trains using block sections containing these switches, for ensuring the correct position of the switches themselves. Of course, all switches in the railway network require the same attention, but in large stations where many lines rely on relatively few platforms, the route management in the vicinity of the platforms may become highly critical.

Before discussing capacity allocation in the following sections, we can identify here three main differences between the two systems:

1. capacity available for allocation in the strategic planning: in air transportation, a fix theoretical capacity is considered until the actual operations; in rail transportation, some methods have been proposed for specifying actual capacity, and in particular for assessing the feasibility of a particular capacity allocation, as described in Section 4.2;

2. the responsibility for capacity allocation in the strategical planning and during the real-time intervention is given to separate stakeholders in the air transportation system and to a single one in the rail transportation system: in air transportation, the airport slot coordinators are responsible for capacity allocation in the strategic planning, and the ASNPs during the real-time intervention; in rail transportation, the infrastructure manager is the only responsible;

3. capacity specification along aircraft and train routes depends on evaluations which are different in nature: in air transportation, sector capacity depends on the number of aircraft that a controller can handle, and the most critical factor is the aircraft entrance flow; in rail transportation, capacity depends on physical constraints, due to train braking distances and the signaling system implemented, and the most critical factor is the number of train services operated.

4 Strategic planning

The main system level activity characterizing the strategic planning in both the air and the rail transportation systems is the capacity allocation to competing applicants. The requests of these applicants depend on their desired timetable. In this framework, the situation is different in the two systems.

In fact, in the air transportation system, the competitions between airlines for obtaining the right of using airport capacity is very strong. As a consequence,
for enduring the avoidance of arbitrariness, the process for allocating capacity is formally defined with a quite high level of detail. Instead, the strategy followed for defining each airline timetable is considered as a competitive advantage, it is hence strictly private information and there is no official rule to be followed. Thus, our analysis will focus only on the process implemented when the desired timetables are disclosed, i.e., the process implemented for allocating airport capacity.

A different situation occurs for railway undertakings where a virtual monopoly often exists for passenger traffic, and transport authorities may have an active role in the timetable definition (see Section 2.2). Hence, in this section, we also report the main trends concerning timetabling. Both freight and high-speed trains are typically not subject to a direct influence of transport authorities. For freight traffic, an early planning of the demand is often not possible. Hence railway undertakings transporting freight take part to the capacity allocation in the strategic planning only for what concerns a portion of their future needs, and make their requests for the remaining capacity shortly before the actual operations. These requests are called *ad hoc requests* and their handling is separately regulated at the European level (European Commission, 2001, 2012b). For high-speed trains, the railway undertakings typically operate on a more strictly commercial basis than for conventional trains, where, to some extent, the public service must be supplied fulfilling certain requirements (for example in terms of train frequency or territorial coverage and cohesion).

### 4.1 Air transportation

In the air transportation system, airlines define their ideal timetable based on commercial evaluations and on constraints imposed by the available fleet and crew. However, these ideal timetables are not always compatible with the constraints imposed by the capacity of the system, mostly by the airport capacities.

In the strategic planning, timetables must respect the airport theoretical capacities, i.e., the declared limit on the maximum number of slots that can be used at each airport per unit of time.

In Europe, a quite elaborated administrative slot allocation process is currently implemented. It follows the rules and principles described in Regulation 95/1993 of the European Commission (1993) and its subsequent amendments (European Commission, 2002b, 2003, 2004b, 2007a, 2008a, 2009c). They are the evolution of a system created by the International Air Transport Association (IATA) in 1947. In this process, the Member States identify the most congested airports and denote them as “coordinated” (“facilitated” if congested only at peak time). These are the airports where an airport coordinator must allocate a slot to an airline, before the airline itself can actually exploit the corresponding facilities (European Commission, 1993). Each coordinated and facilitated airport has its own coordinator who fulfills multiple *tasks* (European Commission, 1993; IATA, 2011):

1. he specifies the declared airport capacity by fixing the number of slots
available per time unit;

2. he guarantees the *grandfather rights*, i.e., the rights of airlines to exploit the slots they have actually used in the preceding equivalent season. In particular, he implements the *use-it-or-lose-it rule* by, on the one side, identifying as slots subject to grandfather rights the ones used for no less than 80% of the time during the previous season; on the other side, considering free from grandfather rights all the other slots;

3. he allocates series of slots to airlines according to the following steps: first he allocates the grandfathered slots (slots on which grandfather rights exist), second he allocates half of the non-grandfathered slots to new entrants (airlines with limited presence at a coordinated airport), third he allocates the remaining slots to unallocated requests according to their subordinate priority (IATA, 2011);

4. he warrants the operational and legal feasibility of subsequent slot exchanges between airlines.

If an airport is neither coordinated nor facilitated, i.e., if it is not congested, the concept of slot is not defined. Airlines do not need to wait for the completion of the slot allocation process for having the right of landing or taking-off at these airports. These airports are often referred as regional airports (Henke et al., 2010).

Airport coordinators’ task 3 constitutes the so-called *primary allocation*. This phase of the slot allocation process is initially pursued by each airport independently from the others, but then it is discussed and adjusted by the representatives of airports and airlines meeting at an IATA conference. Despite these discussions and adjustments, the allocation resulting from the IATA conference may not allow the smooth execution of the timetable desired by airlines. In fact, by allocating slots at airports separately, it may happen that an airline wishing to schedule a flight from, for example, Paris Charles De Gaulle (France) to London Heathrow (UK), receives a slot in Paris Charles De Gaulle at 9:00 am, and at London Heathrow at 9:30 am. These slots clearly do not respond to the airline desires, but no rule is implemented in the current slot allocation process for ensuring the coherence between allocated slots and ideal timetables. After the IATA conference, bilateral negotiations between airlines may continue, leading to possible slot exchanges. These bilateral negotiation are commonly referred to as *secondary trading* (European Commission, 2008a). Through this secondary trading, airlines try to obtain slots matching their desired ones, for finally publishing a timetable which is as close as possible to their ideal one. Any slot exchange resulting from the secondary trading must be explicitly approved by the concerned airport coordinators (task 4). Currently, a well defined legislative basis does not exist for managing the secondary trading of airport slots in Europe. However, the European Commission is assessing some modifications to the current regulation for making official the secondary trading. Currently, slot exchanges with monetary counterparts take place quite regularly at some airports (namely at Heathrow and Gatwick airports, UK).
Airports specify theoretical capacity
Airports declare theoretical capacity and airline request slots
Secondary trading
Airlines publish the timetable
Minor scheduling adjustments
Starting of the season

Figure 8: Schematic representation of the slot allocation process at coordinated airports, i.e., the process of scarce capacity allocation in the air transportation system.

This slot allocation process is schematized in Figure 8 and it is actually rather unsatisfactory. The reason are various: unsatisfied or unaccommodated demand, late return of unwanted slots, flights operated off slot times (off-slot)\textsuperscript{2}, and failure to operate allocated slots (no shows)\textsuperscript{3}. These last three issues may be partly due to the difference between the received slots and the ones desired by airlines: airlines may accept to operate on rather tight timetables, reducing buffer times, for exploiting their slots; this may translate into frequent impossibility of actually operating flights as scheduled. The impact of flights operated off-slot has been measured to be very high in terms of delays (Katsaros and Psaraki, 2012).

According to the Airport Council International Europe (2004), a non negligible percentage of slots allocated after the IATA conference does not translate into actually operated flights. For example, this percentage for the 2002 summer season was 21.4\%, 17.6\% and 2.7\% for Milan Malpensa (Italy), Paris Charles De Gaulle (France) and London Heathrow (UK), respectively. All airports had refused slots to airlines during the slot allocation process due to capacity constraints. By 2030, the demand is expected to exceed airport capacities of about 2.3 million flights (Eurocontrol Experimental Centre, 2008). Part of this increased demand will be absorbed by regional airports (Henke et al., 2010), even

\textsuperscript{2}Off-slots may be due to the tightness of the slots allocated to an airline, which may thus not be able to respect them in case of even a minor disruption. If an airline operates off-slot intentionally and repeatedly then it may be fined by the airport.

\textsuperscript{3}No shows are due to similar reasons as off-slots, and they happen in case of major disruptions causing a flight cancellation. As for off-slots, a fine may follow intentional and repetitive no shows.
if a great part of it will still be directed to main airports. This will strengthen the impact of the unsatisfactory airport capacity allocation. To limit this impact, the slot allocation process must undergo some major revisions.

These major revisions are envisaged by several studies. Some of them (DotEcon Ltd, 2001; Kösters, 2007; NERA Economic Consulting, 2004) underline the necessity of allocating coherently the slots at origin and destination airports of each flight, according to the desired airlines’ timetable. In fact, in the adjustment which follows the primary allocation, during the IATA conference, some inefficiencies may arise in terms of slots allocated either incoherently at couples of airports, or unsatisfactorily from the airlines perspective. These inefficiencies cannot always be corrected in the secondary trading. However, the great majority of the studies either on the primary allocation (DotEcon Ltd, 2001, 2006; Maldoom, 2003; Sentance, 2003; Starkie, 1998; Verhoef, 2010) or on the secondary trading (de Wit and Burghouwt, 2007; DotEcon Ltd, 2001; Holt et al., 2007; Mott MacDonald Limited, 2006; NERA Economic Consulting, 2004; UK Civil Aviation Authority, 2001) does not deal with this necessity. Most of these studies stress the need of an improved slot allocation process for maximally exploiting capacity, enhancing competition and reducing the existing barriers to entry. To this aim, they try to quantify the inefficiency (for example in terms of late return of unwanted slots) of the current process and its improvement potentials.

The remaining papers focus on the qualitative investigation of the opportunity of implementing market-based mechanisms in Europe, following the current practice at some US airports.

The greatest part of the literature focuses on each airport independently, as it happens in the current slot allocation process. The first authors considering the interdependence among airports are Rassenti et al. (1982). They propose a combinatorial auction of slots. However, the main focus of their paper is on the efficiency and robustness of the auction design in terms of demand revelation. More recently, Castelli et al. (2012) and Pellegrini et al. (2012a,b) have considered the interdependency of slots at different airports, respectively for the primary allocation and the secondary trading. These two last papers are the only ones proposing actual methods for dealing with the slot allocation process quantifying the impact of these methods, even if on random generated instances. A quantitative analysis on real instances is proposed by Zografos et al. (2012), who propose an optimization-based model implementing the current IATA rules and the existing operational constraints at an individual airport. The usefulness of this model consists in the elimination of the inefficiencies due to the complexity of the problem to be solved by the slot coordinators.

The implementation of any market mechanism for airport slot trading is based on the hypothesis that airlines can actually associate a monetary value to each slot which it might possibly buy or sell. Recently, Babic and Kalic (2012) proposed a formal model for assessing slot values in terms of potential revenue, aircraft availability, operating costs and period of return on the investment.

Neither the administrative nor the market-based mechanisms envisaged so far appear satisfactory, as discussed by Barnhart et al. (2012). According to the
authors, anyway, the evaluation of the different mechanisms is often impractical due to the difficulty of forecasting airline responses to specific slot allocation processes.

Once the timetable is published, the only system level activity to be performed before actually operating it is the definition of the flight plan. A flight plan is the specified information provided to air traffic control (ATC) units, relative to an intended flight or portion of a flight of an aircraft (International Civil Aviation Organization, 1996). It includes basic information such as departure and arrival locations, estimated duration, alternative airports in case of bad weather, type of flight (whether following instrument flight rules or visual flight rules), information on the pilot, number of people on board and information on the aircraft. Airlines choose the flight plan based on several criteria, including winds and other weather conditions, fuel usage, en-route turbulence predictions, safety constraints, etc. The flight plan is submitted to the Central Flow Management Unit of EUROCONTROL (CFMU) between five days and one hour before the flight departure. The system checks the flight plans against the airspace structure, e.g., in terms of sequence of way-points to be met. Once the flight plan is accepted, it is distributed to the ATC units which will be concerned with the flight.

4.2 Rail transportation

Even if with some possible intervention by transport authorities, in the rail transportation system the railway undertakings construct their own timetable and then request the necessary capacity to the infrastructure manager, as in the air transportation system. Figure 9 schematizes the capacity allocation process for national and international train paths.

For what national train paths are concerned (Figure 9(a)), the infrastructure manager must declare the characteristics of the network in the network statement, at least four months before the deadline for requests for infrastructure capacity. This deadline must be no more than twelve months in advance of the entry into force of the timetable. The capacity allocation concerning national train paths is to be provided no later than four months after the deadline for path requests. European regulations (European Commission, 2001, 2012b) formalize the capacity allocation principles. According to these regulation, in the network statement, the infrastructure manager indicates the structure of the network, the conditions of access to it, the charging principles and tariffs, the main criteria for capacity allocation and the likely capacity requirements for maintenance. Based on this information, the capacity allocation is done once or twice (e.g., in the UK) per calendar year following the above mentioned schedule, and framework agreements may be put in place between the infrastructure manager and the railway undertakings for ensuring capacity allocation for a longer period, typically of five years. Once capacity is allocated, no exchanges are possible between railway undertakings.

For what international train paths are concerned (Figure 9(b)), common procedures must be followed at a European level, and the European Commission
Railway undertakings request train paths. Infrastructure manager proposes a capacity allocation. Coordination between railway undertakings and infrastructure manager. Possible additional path requests. Entry into force of the timetable.

Infrastructure manager publishes the network statement.

(a) National train paths

(b) International train paths

Figure 9: Schematic representation of the process of capacity allocation for national and international train paths in the rail transportation system. In Figure (a), k is the deadline for path requests, which is not imposed by regulations.

must be informed of these procedures (European Commission, 2001, 2012b). In particular, common procedures must be followed by the infrastructure managers whose decisions have an impact on the infrastructure managers of other countries. For having a coordinated management of international train paths, a number of European infrastructure management and railway undertakings formed an organization named RailNetEurope (RNE) in 2004. This organization deals with the capacity allocation at the international level, defining the rules and the procedures implemented in accordance with the European regulations. The path design procedures, in particular, are based on a software tool named Pathfinder\(^4\) which identifies the usable train paths. The schedule to be followed for the request submission and capacity allocation imposes deadlines that precede the ones imposed for national capacity allocation. The great majority of international train paths concern high-speed trains, hence using at least in part a dedicated infrastructure.

\(^4\)http://www.pfndr.org
For both national and international train paths, requests for the use of capacity may only be refused if there are viable alternatives. These alternatives must allow the railway undertaking to operate the freight or passenger service concerned on the same or alternative routes under economically acceptable conditions. When a request is refused, the infrastructure manager must justify the refusal decision and indicate the alternatives. Hence, when the infrastructure manager encounters conflicts between different requests, it shall attempt to meet all requests as far as possible. If no viable alternative is available and it is not possible to accommodate all requests, the excluded railway undertakings may complain to the regulatory body established by the Member State. The regulatory body examines the cases and takes actions, where pertinent, to ensure that an appropriate part of the capacity is granted to the complaining railway undertakings. Moreover, the regulatory body of each Member State may intervene in the capacity allocation process to guarantee that capacity is allocated to priority services, for public-service requirements or for promoting the development of rail freight. The concerned Member State may grant the infrastructure manager some compensation for the consequent loss in revenue. These capacity allocation constraints and impositions clearly exemplify how the European Commission (European Commission, 2012b) urges the increase of competition and efficiency in the railway system.

Due to the complexity of capacity specification (see Section 3.2.1), typically the infrastructure managers rely on simulation tools for verifying the feasibility of different scenarios. Barber et al. (2007) propose a review of the main simulation tools which are used by the infrastructure managers in Europe.

The field of capacity allocation to multiple actors in railway transportation is extremely new, coming as a consequence of unbundling and market liberalization (see Section 2.2). For this reason, and due to the fact that regulations are not well formalized yet, the academic world has not focused much on possible market mechanisms for effectively allocating scarce capacity to competing operators. Instead, a lot of research papers focus on the definition of timetable. This is not exactly a task which is accomplished at system level. Nonetheless, we shortly report the major research directions which have been investigated, since in several countries the timetable is still designed considering a virtual monopoly of the major railway undertaking over the infrastructure capacity.

Many authors tackled the problem of optimizing the timetable. The methods proposed are summarized in several surveys (Assad, 1980; Caprara et al., 2007; Cordeau et al., 1998; Haghani, 1987; Hansen, 2006; Huisman et al., 2005; Liebchen et al., 2008; Lusby et al., 2011; Törnquist, 2006). Typically, these methods seek for the optimal timetable at a macroscopic level, i.e., considering only lines and neglecting detailed constraints at stations and junctions. These constraints are considered in a subsequent phase, tackling the so called train routing problem where the timetable is given. If this problem results infeasible, then the timetable needs to be redefined. Zwaneveld et al. (1996, 2001) and Burkolter (2005) describe different procedures in which the timetabling and the train routing problem are solved iteratively, until a feasible solution to both is found. Schleachte et al. (2011) propose an algorithm for adding trains to an
existing timetable on an aggregated model of a microscopic infrastructure representation. The authors call this aggregation “macrotization”. Subsequently, the infrastructure representation is reconsidered in microscopic terms for assessing the feasibility of the timetable produced and for possibly repairing it.

Two main branches of research emerge for tackling the timetabling problem. In the first branch, the timetabling problem is considered to be cyclical: as mentioned in Section 3.2.1, the same train schedules are repeated every 30 or 60 minutes, for example. For modeling this problem, typically authors formulate a variant of the Periodic Event Scheduling Problem (PESP). Several papers propose exact and heuristic algorithms for tackling the PESP in the railway framework (Kroon and Peeters, 2003; Liebchen and Möhring, 2007; Nachtigall, 1996; Nachtigall and Voget, 1996, 1997; Peeters, 2003). In the second research branch, the timetabling problem is planned for a 24 hour period, without considering cycles. In this case, the problem is typically formulated as a multicommodity flow problem (Borndörfer and Schlechte, 2007, 2008; Caprara et al., 2001, 2002, 2006). Further papers deal with a different but related problem, i.e., the problem of inserting additional trains in an existing timetable (Burdett and Kozan, 2009; Cacchiani et al., 2010; Flier et al., 2009; Ingolotti et al., 2004; Lova et al., 2007; Tornos et al., 2008), mostly for scheduling additional freight trains. In this case, the timetable already established for passenger trains is not modifiable, and freight train operators indicate their requests to the infrastructure manager in terms of an ideal timetable. The infrastructure manager accommodates new trains trying to respect as much as possible this ideal timetable.

As discussed by Lusby et al. (2011), different objective functions have been considered in the literature. Mainly, the timetable is optimized according to three alternative perspectives: the first one, adopted for example by Wong et al. (2008), maximizes passenger satisfaction by minimizing factors as the waiting time; the second one, adopted for example by Fischetti et al. (2009), maximizes the timetable robustness, as defined in the following; the third one, adopted for example by Carey and Lockwood (1995) and Caprara et al. (2002), minimizes the profit loss due to the discrepancy of the allocated train paths with respect to the ideal ones.

Part of the literature on rail timetabling is devoted to assessment methods. The most important characteristics that a timetable must have are the following (Vromans, 2005):

- reliability: it must be possible to perform the trains scheduled for a specified, uninterrupted, period of time;
- robustness: it must be possible to perform the trains scheduled under difficult circumstances without major modifications in the timetable;
- stability: after a disturbance it must be possible to return to normal operations with limited time and effort.

Assessing reliability, robustness and stability in analytical terms is extremely complex, if ever possible for large networks. The research in this context fo-
cused on optimization and simulation techniques and tools (Goverde, 2005; Kuni-
matsu et al., 2009; Lindner, 2011; Medeossi, 2009; Medeossi et al., 2011; Soto y Koelemeijer et al., 2000; Vromans, 2005; Yuan, 2006). Different algorithmic approaches focus on a single characteristic among the mentioned ones: for example, random perturbation may be used to measure the robustness of the system (Odijk, 1999).

For what concerns the train routing problem, relevant research has been proposed along two main directions. The train routing problem has been modeled as a graph coloring problem (Billionnet, 2003; De Luca Cardillo and Mione, 1998) and as a set packing problem (Gandibleux et al., 2004; Lusby, 2008; Velásquez et al., 2005; Zwaneveld et al., 1996, 2001). It is not possible to identify the best approach among these, as they consider different frameworks. For example, Billionnet (2003) focus on the problem of assigning trains to platforms when the train arrival time at the station is not modifiable, while Zwaneveld et al. (2001) consider the possibility of having the train shunted in a parking area outside the station itself.

4.3 Main similarities and differences

The nature of the problem which must be solved for allocating infrastructure capacity in the strategic planning is equivalent in the air and rail transportation systems: scarce capacity must be allocated to several operators that are in competition with each other. However, in the analysis of the procedures implemented, some major differences emerge. The most important concern:

1. the part of infrastructure which is considered to be relevant for capacity allocation: in air transportation, it concerns only airports; in rail transportation, it concerns the whole network;

2. the characteristics of the requested operations that are considered for capacity allocation: in air transportation, requested flights are considered only for what concerns airports, airports are managed separately from one another, and so airlines are in charge of matching the allocated capacity with their desired timetable; in rail transportation, train movements are considered entirely;

3. the level of time detail considered for capacity allocation: in air transportation, capacity is allocated by assigning the right of landing or taking-off at an airport in an interval of some minutes to a number of airline, and the specific order of their use of the infrastructure is not indicated; in rail transportation, capacity is allocated by indicating the exact order in which trains may use the infrastructure, with a precision of fifteen seconds to one minute, typically;

4. the level of analysis implemented: in air transportation, a one-level model is sufficient, since capacity is allocated at individual airports where the
runway system is the constraining resource; in rail transportation, a two-
level model is necessary for dealing with capacity allocation at the network
(macroscopic) and at the junction level (microscopic);

5. the information used when allocating capacity: in air transportation, ca-
pacity is allocated to airlines without considering the flight and type of
aircraft for which they are requesting slots; in rail transportation, the in-
frastructure manager takes into account many details on the requested
paths, as the rolling-stock that will be used.

5 Real-time intervention

After timetable definition and capacity allocation, the operations actually take
place, and the main system-level activity is traffic management. It includes all
the activities which are necessary for ensuring the effective and safe respect of
the timetable. These activities may consist in the straightforward application
of the decisions made in the strategic planning, or in the solution of emergent
problems due to unplanned events. These events may be either internal, due
for example to the absence of crew members that impose a rescheduling, or
external, due for example to bad weather conditions. When such events occur,
a real-time intervention may be undertaken for re-allocating capacity to smooth
operations.

A fundamental difference exists between the air traffic management per-
formed by ANSPs and the rail traffic management performed by dispatchers.

ANSP operators are in charge of ensuring safety in the air operations. Particu-
larly for what concerns en-route sectors, they deal with a theoretically very
high capacity infrastructure, since a very high number of aircraft could poten-
tially fly in a sector at the same time, if we considered only physical constraints.
However, ANSPs are responsible of guaranteeing safety: for being able to give
this guarantee, the capacity is artificially restricted to what a person can control,
i.e., according to human factors and ergonomic criteria.

In the rail transportation system, safety of train movements is ensured by
the signaling system (Section 3.2) in the large majority of possible situations,
and dispatchers are in charge of smoothing the operations in case of unex-
pected events. In fact, in absence of unexpected events, the train drivers follow
the timetable planned, which details the suitable route and speed for avoiding
conflicts; the signaling system guarantees that no safety issues emerge. If an
unexpected event occurs, then the worst case scenario is a major delay propa-
gation, with a consequent economic loss. The role of the dispatcher consists in
limiting as much as possible this delay propagation and he is relieved of almost
all safety responsibilities.

5.1 Air transportation

ATM is the dynamic, integrated management of air traffic and airspace – safely,
economically and efficiently – through the provision of facilities and seamless
services in collaboration with all parties (International Civil Aviation Organization, 2005).

ATM performance is measured through a number of indicators and binding targets on the key performance areas of safety, environment, capacity and cost-efficiency (European Commission, 2011b), while the required safety levels must be fully ensured. The maximization of the level of performance is pursued through different types of control actions (Ball et al., 2007; Vossen et al., 2012):

1. **Ground holding.** It consists in delaying the departure of a flight to avoid overloading a system element, be it either an en-route sector or an airport. In the practice, the ground holding is assigned only to flights departing from airports in the vicinity of the overloaded element, and its fair distribution across airports is based mostly on the expertise of air traffic controllers and on the available traffic forecasts. At each airport, the ground holding is imposed on a first-planned-first-served policy, i.e., aircraft are allowed to depart in the order that would have been followed in absence of any restriction\(^5\). In Europe, the ground holding is assigned in a centralized fashion by the CFMU of EUROCONTROL, by imposing departure time slots (ATFM slots) through the Computer Assisted Slot Allocation (CASA) (Eurocontrol, 2013);

2. **Airborne speed control and airborne holding.** They aim at the same objective as ground holding. They are implemented by adjusting the time at which aircraft travel. Speed control consists in slowing down or speeding up aircraft. Airborne holding typically consists in having aircraft fly in oval-shaped patterns.

3. **Route adjustments.** They are typically implemented in the flight plan verification phase, but they can be decided also at a later moment. Sometimes, standard reroute strategies may be specified. For example, all flights originally filed along one airspace path may be rerouted along a different path.

4. **Arrival sequencing.** The sequencing of flights can be very important as the maximum arrival rates into airports depend on the sequence and mix of aircraft types.

Ball et al. (2007) and Vossen et al. (2012) add two more control actions to this list: **flight cancellations** and **way-point flow restrictions**. We prefer to mention these actions separately for different reasons. For what concerns flight cancellation, we do not include it in the list since responsibility for flight cancellations always rests with the individual airline, unless of course extreme measures are taken, such as the closing of the departure airport. In any case, it cannot be considered a system level activity. For what concerns way-point flow restriction, we do not include it in the list since the subject of our analysis is the European air transportation system, and this control action is almost exclusively

\(^5\)Despite the well known sub-optimality of this policy, it is perceived as fair by airspace users. A more effective and as fair-perceived policy has not been identified yet.
implemented in the US: regional air traffic service providers may impose such restrictions to ensure that the flow of aircraft into a region of airspace is kept at a safe level. When such a restriction is put in place, the adjacent “upstream” regional air traffic service provider has responsibility for maintaining a traffic flow at or below the restricted level. This can be done in a variety of ways, including the use of airborne holding, rerouting of some traffic and issuing similar flow restrictions on flights further upstream. In this way, it is possible for a flow restriction to propagate through much of the airspace system, possibly leading to ground holds at airports of origin. The overall effect of way-point flow restrictions is typically not assessed in the practice, since it hardly ever achieves relevant proportions.

In the literature, the first problem which has been tackled in the context of ATM is the ground holding problem. It consists in allocating ground holding delays to a set of flights, trying to minimize an aggregated cost function. This problem has been tackled with both exact and heuristic algorithms, considering either deterministic or stochastic demand, and different levels of complexity of the network (Andreatta and Romanin-Jacur, 1987; Dell’Olmo and Lulli, 2003; Hoffman and Ball, 2000; Odoni, 1987; Richetta and Odoni, 1993; Terrab and Odoni, 1993). Recently, a major issue being investigated is the possibility of implementing a market mechanism for allowing airlines to buy and sell ground holding delay (Castelli et al., 2011; Sherali et al., 2011).

The evolution of this problem, which is currently object of most studies, is the air traffic flow management problem, in which both airborne speed control, airborne holding and route adjustments are possible. Lulli and Odoni (2007) give the following definition: given an airspace system, consisting of a set of airports, airways, and sectors, each with its own capacity for each time period, over a defined time horizon, and given a schedule of flights through the airspace system during this horizon, the air traffic flow management problem consists in assigning ground and airborne delays to the flights in a way that satisfies all the capacity constraints while minimizing a function of the cost of the total delay. This problem is recognized to be very critical in Europe: the density of flights in the sky is often very high, and hence en-route congestion is a major constraint to the increase of air traffic. As for the ground holding problem, several algorithms have been proposed for solving the air traffic flow management problem, tackling networks of different complexities, and adopting either exact or heuristic strategies (Bertsimas et al., 2011; Bertsimas and Patterson, 1998, 2000; Helme, 1992; Lindsay et al., 1993; Lulli and Odoni, 2007; Sherali et al., 2003, 2006). Most of these algorithms are based, at least in some part, on integer programming models. With few exceptions (Bertsimas et al., 2011), they consider fixed travel time between origins and destinations. This is not considered to be an issue in this framework, since the impact of the speed variation that can be imposed by controllers is typically minor. When speed control is considered, a minimum and a maximum time are considered for aircraft to cross sectors, but no speed variation dynamics are modeled. By setting a high maximum time on the sectors corresponding to the terminal airspace, such a model captures also the implementation of airborne holding. Moreover, the al-
algorithms proposed consider aircraft rerouting in terms of sectors crossed: they neglect the specification of airways and way-points. Finally, as in the timetable definition, time indications are given considering some minute time intervals.

Concerning arrival sequencing, Soomer and Franx (2008) and Castelli et al. (2011) propose market mechanisms in which airlines can trade on arrival slots. Through this trade, arrival sequence may reflect the cost associated by airlines to delays due to sequencing itself. The two papers propose, respectively, a local search algorithm and two distributed approaches that allow airlines not to disclose information on costs.

D’Ariano et al. (2010, 2012) propose a different approach for routing and sequencing aircraft in the vicinity of airports in real-time. This approach models the problem as a job shop scheduling with additional real-world constraints, proposing a parallelism between the real-time allocation of airport capacity and the real-time allocation of railway junction capacity, which is the focus of the work of the same authors in the context of the rail transportation system (Corman et al., 2009; D’Ariano, 2009; D’Ariano et al., 2008; D’Ariano and Pranzo, 2009; Pranzo et al., 2005), as discussed in Section 5.2.

The ATM processes which are currently implemented are not considered as the most urgent source of inefficiency of the system. Hence, most research in this field is academic rather than commissioned by institutional stakeholders. What is mostly investigated in the institutional studies is the technological development for what concerns both aircraft and air traffic control centers (Eurocontrol Experimental Centre, 2009) and the reorganization of airspace design (Eurocontrol Experimental Centre, 2011).

In the practice, control actions are decided in advance with respect to actual operations. In particular, when a perturbing situation, such as bad weather condition, is expected, possible ATM control actions are evaluated for selecting the specific ones to implement. This is typically done with a few-hour advance. What is actually performed in real-time is the conflict detection and resolution: potential conflicts between different aircraft trajectories have to be detected and tackled for ensuring safety (European Commission, 2011a). Two types of procedures may be implemented to this aim: Visual Flight Rules (VFR), in which the pilot is in charge of visually monitoring traffic in his vicinity; Instrument Flight Rules (IFR), in which an ATC is in charge of monitoring traffic through radars. Nowadays, the IFR are operated in the majority of the situations. These rules are operated through resolution advisories (RAs) given to the pilots encountering a potential conflict. Although in the academic literature conflict resolution rules based on aircraft speed variations have been investigated (Archambault, 2004; Cobano et al., 2012), in the practice, RAs consist in either climbing or descending maneuvers. Among those ensuring safety, the selected RA is the one minimizing the disruption with respect to the original flight plan (International Civil Aviation Organization, 2006): no system level efficiency assessments are made in this framework. Due to this lack of system level assessment, we do not further focus on conflict detection and resolution issues in this paper, and we refer the interested reader to Kuchar and Yang (2000) and Aoude (2011) for a detailed review of existing models and
development trends.

5.2 Rail transportation

As reported in Section 3.2.2, in the rail transportation system, the capacity re-allocation during the real-time intervention is ensured by the infrastructure manager, and it is performed by dispatchers. Each dispatcher is responsible for a well defined part of the network, and he must communicate to neighboring control centers any relevant pieces of information. Dispatchers typically base their decisions on the indication of software tools that differ from country to country. For example, in the Netherlands, the software tool used by dispatchers is the ARI system, which reserves routes to trains based on timetable orders and train describer data (D’Ariano et al., 2008). However, dispatchers are often forced to take last minute decisions autonomously. The main decisions which the dispatcher makes for re-allocating capacity are: changing the run order at critical locations and rerouting trains. For changing the run order, dispatchers can either temporarily stop some trains through the signaling system, or suggest suitable speed adjustment actions to the drivers.

In the scientific literature, the main issue tackled in the framework of the real-time intervention is the real-time solution of unforeseen conflicts in the network (real-time traffic management problem). The problem to be tackled is somehow similar to the train routing problem discussed in Section 4.2 for the timetable design: trains travel along the same network and they must compete for exploiting the scarce infrastructure capacity. A critical difference is the degree of freedom available. In the timetable design, train routes and schedules are an input only as an ideal timetable. If this ideal timetable cannot be accommodated, any deviation is, at least in principle, possible. In the real-time traffic management problem, instead, trains are already present in the network when capacity issues are to be solved, and they already traveled part of their routes. The spectrum of possible decisions is, in this case, much smaller. A further difference between the two problems consists in the fact that, when designing a timetable, train run-time along each track section is calculated according to the assumption of undisturbed operations (green signal aspects); when managing operations, instead, the run-time must consider disturbed operations (restrictive signal aspects) and brakes and accelerations may have to be added. Moreover, the objectives of the timetable design and of the real-time traffic management problems are typically quite different: on the one hand, for example, the maximization of the robustness is a minor criterion in a real-time intervention framework; on the other hand, the minimization of delay propagation is meaningless in the strategic planning framework, when train schedules have not been fixed yet. Finally, the solution of the real-time traffic management problem must be done extremely quickly.

Several papers propose efficient algorithms for tackling the real-time traffic management problem. They can be divided into two main families. On the one hand, some algorithms seek for the best train reordering and rerouting taking into account the speed variation dynamics, i.e., the realistic brake and
acceleration times and distances. These algorithms are named variable-speed algorithms. On the other hand, in the so called fixed-speed algorithms, the best solution is searched neglecting any speed dynamics: trains are supposed to need neither space nor time for passing from the maximum speed allowed on the line to the complete stop, and viceversa. This neglect clearly implies a possible large gap between the evaluation of a solution made by a fixed-speed algorithm and in reality. Measuring this gap is a very hard task in itself. Rodríguez (2007) reports some experimental evidence on the comparison between the solutions returned by an algorithm when considering either its variable or its fixed-speed implementation. The specific variable-speed algorithms uses an approximation of speed variation dynamics rather than their exact computation. The author assesses these solutions in simulation. Here, the increase of realism achieved through the variable-speed algorithm does not pay, due to the huge additional computational effort for calculating speed variation dynamics and to the consequent limited effort available for the search space exploration. The general validity of this results has not been proven yet, but the number of papers considering fixed-speed algorithms present in the literature, compared to the number of those considering variable-speed ones, shows that the community is quite unanimous on this validity.

Including both fixed and variable-speed algorithms, the research direction which has had the greatest follow-up implies the definition of an alternative graph on which to model a job shop scheduling problem. Since 2007, several papers have proposed algorithms for tackling this type of model (Corman et al., 2010, 2012a; D’Ariano, 2009; D’Ariano et al., 2008, 2007a,b; Mazzarello and Ottaviani, 2007). These papers consider problem variants where train rerouting either is or is not allowed, either a single or multiple objective functions are considered, either exact or heuristic solutions are returned. Further variants of the problem which have been studied in this research stream consider the solution of the real-time traffic management problem over a long time horizon (D’Ariano and Pranzo, 2009), across multiple dispatching areas (Corman et al., 2012b), and implementing the green wave policy (Corman et al., 2009). The green wave policy consists in possibly increasing train stop time at station for having trains always running at their scheduled speed: the train always encounters green signals along its way within stations. This policy is the equivalent of the assignment of ground holding delay in the air transportation system. All the algorithms proposed earlier than the one by D’Ariano et al. (2008) are fixed-speed, but the one proposed by D’Ariano et al. (2007b) who present a heuristic iterative procedure for taking into account speed variation dynamics.

Some other approaches have been tested: Rodriguez (2007) and Rodriguez and Kermd (1998) consider constraint programming techniques for both fixed-speed and approximated variable-speed algorithms; Törnquist (2007), Törnquist Krasemann (2012) and Törnquist and Persson (2007) propose heuristic and exact fixed-speed algorithms for tackling the real-time traffic management problem in a very large network including several stations and junctions.

The difference between fixed and variable-speed algorithm is overcome by some papers (Caimi et al., 2011, 2012), which consider a variant of the prob-
lem in which trains cannot be delayed within the infrastructure under analysis, and they hence always run at the scheduled speed. This variant imposes the definition of coordination mechanisms between confining control areas. In some sense, it corresponds to the implementation of the green wave policy.

Finally, a recent paper (Lusby et al., 2012) proposes an iterative heuristic variable-speed algorithm based on a set packing formulation. For having the tractability of realistic size instances, the time horizon needs to be discretized with a step of 15 seconds. The authors recognize that this actually represents a constraint which artificially limits the control area capacity, but they consider this limit acceptable. Initially, only one route is considered for each train and, if no feasible solution exists, further routes are iteratively taken into account.

5.3 Main similarities and differences

Traffic management in both the air and the rail transportation systems consists in tackling in real-time issues emerging from the lack of capacity; the principal options available for tackling these issues is the modification of the ordering and the routing of trains and aircraft. Despite this common framework, important differences exist. The main differences concern:

1. the responsibility of controllers and dispatchers, as mentioned at the beginning of this section: safety in air transportation (SESAR Consortium, 2008), and efficiency in rail transportation (Vromans, 2005);

2. the consideration of speed variation dynamics: in air transportation, they are considered with strong approximations (Bertsimas et al., 2011), with supposed negligible impact on the validity of the solutions; in rail transportation, they are recognized to be very important, even if they may not deal with, due to computational constraints (D’Ariano, 2008);

3. the level of detail in the route definition: in air transportation, routes are described in terms of sectors crossed (Bertsimas et al., 2011); in rail transportation, routes may consist in the complete sequence of track-circuits which a train uses (Goverde and Meng, 2011);

4. the level of detail in the time consideration: in air transportation, time is discretized using a step of few minutes (SESAR Consortium, 2007); in rail transportation, the time precision to few seconds is considered extremely important (Goverde and Meng, 2011).

6 Conclusion

Both the air and the rail transportation systems are characterized by the presence of different stakeholders with competing objectives, which either manage or exploit the scarce infrastructure capacity. The existing processes for coordinating the activity of these stakeholders are the results of a rather long history of separate national developments. Recently, the European Commission has
envisaged a common future for these systems. This common future consists in abandoning the current fragmentation of the systems for moving toward actual international air and rail transportation systems, involving all Member States. This unification and homogenization is foreseen to have a major positive impact on the development of the European economy.

In this paper, we proposed an analysis of the main processes implemented in the two systems concerning capacity exploitation, for underlying similarities and differences. In this analysis, it emerges quite clearly that several steps have been made so far toward an homogenization of the systems at a European level, and that others still need to be made. In particular, the air transportation system and the rail transportation system for what concerns high-speed international train paths appear more homogeneous than national rail transportation. A reason for this difference may be the more recent development of the two systems, at least on a large scale. Thus, the rather modern legislation aiming at unified European transportation systems has had less historical and institutional obstacles to overcome. The higher homogeneity is of course to be ascribed also to the fact that a high level of coordination between national authorities must necessarily be ensured for making operations possible.

Besides the movement toward European systems, what clearly emerges in this analysis is the need for an increase of efficiency of the systems. The search for such an increase involves practitioners, institutions and academia. Due to historical evolution and technological development, the processes implemented for managing the two systems still present some important differences. However, some similarities emerge. The analysis of these differences and similarities may help in the common evolution of the air and the rail transportation systems.

In particular, one of the most critical issues in both systems is the capacity allocation in the strategic planning. It is at this time that multiple competing stakeholders ask for the right of using the infrastructure capacity in specific time periods. Capacity is allocated to these stakeholders in quite different ways in the air and the rail transportation systems, mostly due to the historical development of the allocation processes and to the number of actual competitors. For the air transportation system, the research branch focused on the study of alternative capacity allocation processes is particularly flourishing, and it includes a large number of both academic and institutional studies. As underlined by Barnhart et al. (2012), the development of creative mechanisms in this framework is one of the highest priority research opportunities over the next five to ten years. For the rail transportation system, instead, capacity allocation is hardly ever considered as a competitive issue. This is due to the virtual monopoly characterizing conventional passenger traffic at national level. Undoubtedly, the actual status of the system does not allow the realistic expectation of a change in this situation in the immediate future. However, the introduction of a suitable capacity allocation mechanism may be a first step in the direction of this change, and a critical mass of studies is certainly necessary before planning an actual capacity allocation mechanism implementation. It is mostly in this framework, in our eyes, that the experience gained in the air transportation system may be of great help in the development of a European rail transportation system.
A further knowledge transfer between the two systems may concern the real-time intervention modeling. In the rail transportation system, decision aid tools are being studied for supporting dispatchers when managing perturbed traffic. The level of detail considered is very high, in the modeling of both the infrastructure and the train dynamics. Even if the research branch focusing on effective approaches in this framework is far from being exhausted also in rail transportation management, some very interesting approaches have already emerged. In the air transportation system, instead, the real-time intervention completely relies on controllers. The main proposals for decision aid approaches base their modeling on an infrastructure simply made of sectors, while maybe modeling airways might help in better managing traffic flows. In this context, the experience gained in the studies concerning the rail transportation system might be used as a starting point.

References


DotEcon Ltd, 2006. Alternative allocation mechanisms for slots created by new airport capacity.


Kösters, D., 2007. Study on the usage of declared capacity at major German airports.


NERA Economic Consulting, 2004. Study to assess the effects of different slot allocation schemes.


UK Civil Aviation Authority, 2001. The implementation of secondary slot trading.


Zografos, K., Salouras, Y., Madas, M., 2012. Dealing with the efficient allocation of scarce resources at congested airports. Transportation Research Part C 21, 244–256.
