Centre for Transport Studies
STOCKHOLM

# Congestion and scarcity in scheduled transport modes

Jan-Eric Nilsson, VTI

CTS Working Paper 2012:25

*Abstract:* This is a draft text for a chapter in **Handbook on Research Methods in Transport Economics and Policy** to be published by Edward Elgar Publishing. It provides an overview of issues related to scarcity in scheduled transport modes with emphasis on railways. The paper separates the scarcity or time-tabling problem into two analytical parts. The first concerns the challenges related with finding an approximate solution to the mathematically challenging NP complete problem. The second generic problem is related to base the solution to this challenge on the operators' value of each departure slot.

*Keywords*: time-tabling, track scarcity, runway scarcity, willingness to pay.

*JEL Codes*: D61, D82, R42

Centre for Transport Studies
SE-100 44 Stockholm
Sweden
www.cts.kth.se

vti

KTH

## 1. Introduction[1]

Congestion is a familiar concept within the road sector, first and foremost since it provides a very hands-on and frequent experience for many drivers. Analytically, congestion is an external effect within the collective of drivers in so far as the marginal driver causes extra time for those already in the system without necessarily taking this into account. This is the basis for a proactive policy towards congestion, designed in order to ascertain an optimal mix of pricing and investment principles to handle the imbalance between supply of, and demand for roads.

The meaning of congestion in modes where scheduling is a prerequisite for operations differs from the definition of congestion in the road sector. There are furthermore both similarities and differences in the way in which congestion manifests itself when comparing different scheduled modes. In the railway sector the presence of high demand means that there are not tracks available to cater for the wishes of all railway operators, or that some must adjust their demand for access in order to cater for the demand from one or more other operators. As a consequence, services may not commence without operators having been allocated a slot long before a train is about to leave. We will characterise this as a scarcity problem which is solved by constructing a time table that can be advertised to travellers well in beforehand.

Railways also encounter a congestion problem since the pre-set schedule may be disrupted, causing delays which can spread through the system. Reasons can be exogenous, bad weather being the most apparent example. But the disruption can also be internal, caused by one or the other agent within the system, be it a train driver, a vehicle with technical problems, delayed infrastructure maintenance or a dispatcher's handling of a situation in a not so diligent way. Moreover, an original disturbance can have substantial consequences for others using the railway network. Irrespective of cause of the initial perturbation, it is necessary to establish priority principles for handling the situation. With or without priority principles, congestion – but not scarcity – may manifest itself in the same way as in the road network, i.e. by way of railway lines clogging up.

[1] This work is funded by Centre for Transport Studies. I am grateful for comments on previous versions of the text from Chris Nash.

The situation is in many dimensions similar in the air transport sector. In particular at major airports, demand for landing and take-off slots may exceed runway capacity. An established schedule may also be disturbed by external or internal events, requiring real-time reshuffling of departures or arrivals. An important difference compared to railways is, however, that the degrees of freedom in both scheduling and in particular in dealing with disturbances are higher. For this reason, we will address scarcity of airport slots but will not deal with airport congestion further.

Also bus and coach services are scheduled. The reason for not dealing with scarcity in this context is that even very strict bus timetables can be adapted to the precise situation at hand without having to grapple with the problems which come up in railways and at airports. There are even more degrees of freedom than in the airline business.

Against this background, the primary purpose of this chapter is to address the question about how to settle time tables in the rail and air industries, i.e. how to give demand from some operators priority over demand from others. While the presence of multiple users is the standard for most airport facilities, this has not been so for railways until recently. As will be demonstrated, reforms of Europe's railway industry, with vertical separation of infrastructure and operations in combination with a gradual increase in the number of operators, have made the two modes similar in this respect. An additional purpose of the chapter is to address some aspects of railway congestion.

The text is organised in the following way. In section 2 the nature of the scarcity and congestion problems will be specified in greater detail. Section 3 identifies the relevant objectives for addressing these problems while section 4 reviews current approaches to handle them. Contrasting section 3 to section 4, section 5 establishes shortcomings of existing procedures while section 6 suggests different means to handle these flaws. Section 7 concludes.

## 2. The nature of the problem

Figure 1 illustrates generic features of the time-tabling problem by way of a (simplified) string diagram for a single-track railway line.[2] This type of diagram is a widely used workhorse for characterising the capacity problem, for the analysis of alternative solutions as well as for compiling the time-table which is finally established. It is commonly referred to as the graphical time-table. While it is not used for scheduling starts and landings at airports its logic carries over to the airline industry.

In these graphs, each railway station is depicted by a horizontal line, a station being a place where trains may meet or overtake each other; here, the figure has four such stations (I- -IV). A horizontal move along the line (on the x-axle) represents a movement in time. A string between two stations therefore describes a train, and the steeper is the string, the faster is the train.

Train A leaves station I at some specific point of time (say 07:00). It stops at station II (say at 07:16), leaves again (at 07:18) and arrives at the terminal station IV (at 07:46). The movement of train A from stations I to IV represents a route comprising a set of consecutive blocks, a block being a section of a railway line that can be occupied by at most one train at a time. On the single-track line in our example this is the stretch between meeting-stations I and II, II and III etc. Train B describes a train going in the other direction. If train B is the same physical rolling stock as train A the two represents two legs of a round trip.

Using these specifications we can establish some features of track capacity supply:

1. When one train makes use of a block, others cannot use it at the same time; there can be zero or one train at a time on each block.

[2] All critical features of the time-tabling problem can be characterised by detailing the problems encountered on a single-track line. Introducing a second track makes the time-tabling problem easier since it provides more degrees of freedom to the planner in view of that all trains on each track can move in one direction only.

2. A block is much longer than the train. Even if demand for access to a common infrastructure emanates from a fairly small number of train operators, a capacity shortage may occur.
3. Once built, supply-adjustments to accommodate higher demand are lumpy, costly and take a long time before becoming operational.

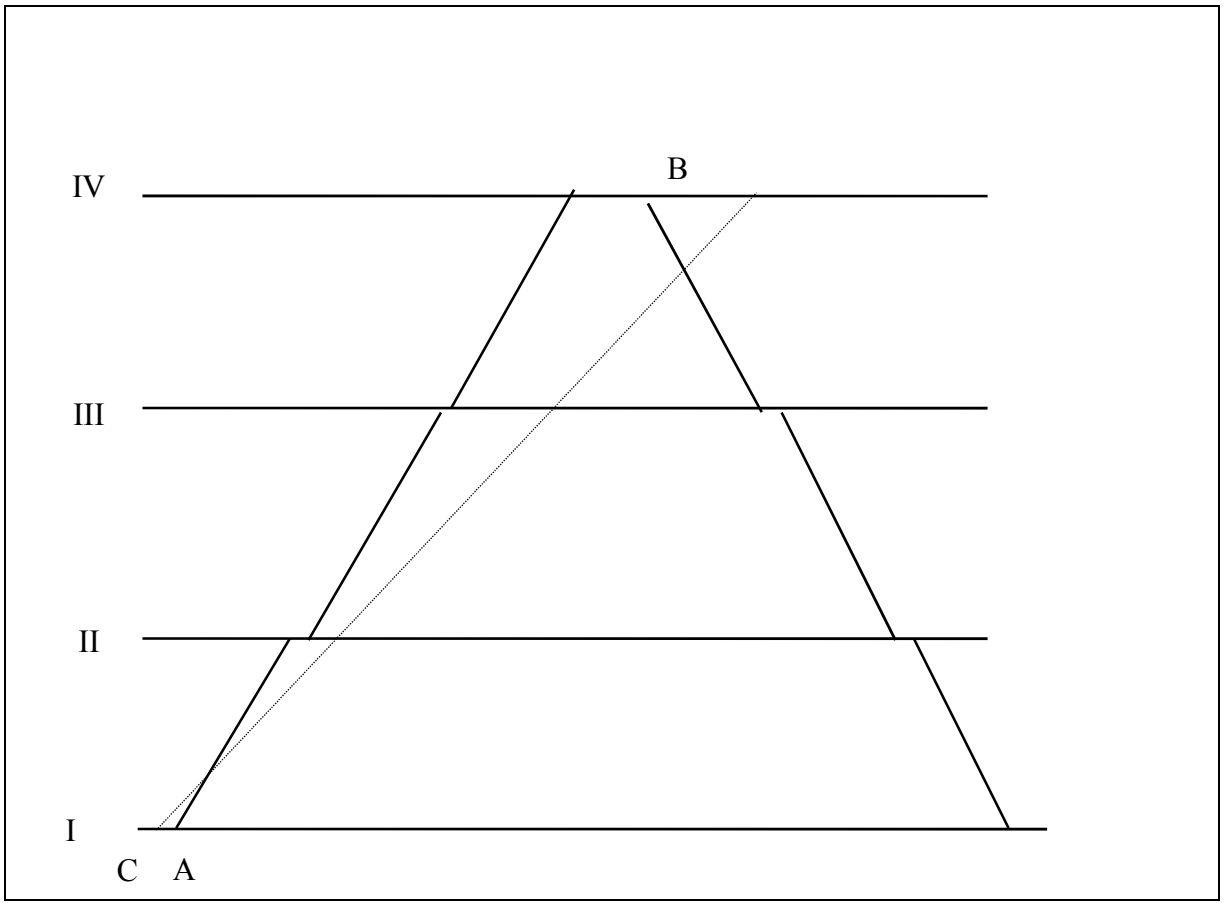


Figure 1. A string diagram depicting the movement of three trains over a certain network.

Demand for track access is derived from the demand by passengers and freight customers to travel or to send freight loads between nodes in the network. The following aspects of track access demand can be observed.

1. Each departure time typically has (a large number of) substitutes, but each of these may have a different economic value: While a passenger service operator prefers to leave at 7:00 it may be feasible, although less valuable, to leave at 6:45 or 7:15 or at some time in-between. Since train operators have to meet customer demand

there are strict boundaries on this flexibility; passengers are simply not interested to leave earlier than, say, 6:30 and a set of freight cars would not be loaded and ready to leave before some given point of time.

2. There may be cyclical variations in demand creating peak periods where available capacity is insufficient and slack periods with under-utilised capacity. In conjunction with the previous point, demand cyclicity can be used to overcome capacity shortages by way of adjusting demand from operators with larger flexibility than others.
3. There may be complementarities in the demand for track access. Going from Here to There could be valuable only if it is possible to proceed to Elsewhere at an appropriate point of time. The allocation process must therefore be designed so that an operator can co-ordinate its services, including return legs where necessary, to achieve efficient use of rolling stock. It must also facilitate coordination between different operations.
4. Services of different operators may also be each other's substitutes. Two firms running the same type of train over a section of the network may value track access differently depending on how close to each other the respective routes over the network are laid.

These demand and supply aspects must be taken into account when the time-tabling problem is to be handled. The presence of an operator of (the slower) train C in the figure illustrates the nature of the type of trade-off which has to be made. Train C wants access in a way which is not compatible with the wishes of the operator(s) of trains A and B; crossing strings outside stations indicate prohibitive conflicts. The question is whether train C or the combination of A and B should be given priority, or indeed if one or more of them could be induced to adjust their demand so that all three can be licensed.

Services provided by an airport comprise four components; the passenger and the freight terminals – often separate buildings – the air traffic control system and the runways. Emphasis is here on runway capacity but the discussion bears over also to other parts of capacity if this is more congested.

While railway capacity may relate to some or all blocks between Origin and Destination, including also scarcity at nodes, focus here is on the situation at the nodes, i.e. the airports. There are, indeed, also en route scarcity problems, but since planes may fly at different altitudes the degrees of freedom are still much higher than on the ground. The en route challenges, which often include flights from different countries, are therefore not further dealt with here, although the scarcity problem does not differ from the problem handled at the nodes and in reality is similar to that in the railway sector.

It is not necessary to make any substantial changes in, or additions to the above description of the supply and demand features of railways in order to account for the situation at airports. As for complementarities in the demand for access, acquiring a slot on congested airport A is relevant only if the airline is allocated a slot also at the destination and if this slot fits in with demand for slots for further departures. The allocation process must therefore be designed so that an operator can co-ordinate (different legs of) its services, and also to facilitate coordination between different operations.

## 3. What are the objectives?

Airport and railway slots are critical to the operation of railway and airline services, since it is not possible to operate trains or to fly without being awarded the relevant slots. The way in which the respective markets function is therefore dictated almost directly by the way in which scarce slots are allocated.

The criteria for prioritising between alternative mechanisms for scheduling infrastructure are the same as when processes for the allocation of scarce economic goods at large: With due respect for equity concerns, efficiency requires that scarce resources are allocated to those customers which value them most. Grether et al. (1989) provide further specification of the policy challenges for a capacity allocation process which strives to establish an efficient and equitable allocation of airport slots. Their discussion is of direct relevance also for railways.

Efficiency requires the market to expand in a way which grants operators with efficient marginal operations (relative to the marginal operations of others) access. Since a slot is a critical resource, those which have relatively high profit[3] opportunities for additional slots should be allocated additional slots by the system. The argument bears over also to a potential entrant. As a corollary, the slots for expanding carriers should come from carriers whose marginal operations are the least efficient. Marginally unprofitable services, and less successful operators, should therefore be induced to exit from the business by the allocation mechanism in use.

Efficiency is also enhanced by postponing demand from operators which have some latitude within which services can be shifted, giving priority to operators with less flexibility. This means that gains from exchange must be achieved through proper coordination of demand.

It has also been established that there exist interdependencies among operations at different railway lines or airports since any train/flight must involve (at least) two terminals. A system of slot allocations must be capable of capturing the efficiency gains and reductions in the overall costs which can result from proper coordination among lines and airports in order to enhance efficiency of the system as a whole.

[3] The issue of profits versus social benefits is further addressed in section 6.4.

It has already been established that the allocation of slots dictates the pattern of market competition at large. The control of slots could therefore provide a key for the development and enforcement of anticompetitive practices. It is therefore reasonable to design the process of allocating slots in a way which provides adequate safeguards against monopoly and/or collusion. In addition, the allocation procedure should be reasonably cheap to administer, including the oversight that is necessary to minimise the risk for collusion etc. This is the transaction cost dimension of the mechanism.

There is furthermore an element of dynamic efficiency or long-run industry growth related to the way in which scarcity is dealt with. If demand for railway or airline services grows over time additional capacity is required. Yet, capacity expansion necessarily absorbs valuable resources. One measure of the need for capacity expansion is the value created by additional slots. If such values, when integrated over time, exceed the cost of expansion it is reason to consider the construction of additional capacity. The allocation procedure should be supportive to this end, i.e. it should produce information about the value of additional slots. In situations where there are absolute barriers to capacity expansion, the need to ascertain efficiency in the allocation of the slots that after all are available is even more acute.

In addition, equity concerns may be relevant. One obvious example from the transport industry is that services to small communities are to be given due consideration by the allocation method. If the political system finds it relevant to account for this type of concern, these objectives should be spelled out in some detail and the organisation in charge of the allocation process should be instructed to handle these concerns.

# 4. How are scarcity and congestion handled today?

The operators' total cost for running trains and airlines have to be covered by revenue from ticket sales to passengers and tariffs for freight customers plus possibly public sector subsidies in one form or another. One component of the operators' total costs is the fees levied for access to infrastructure. The higher the level of these fees, the lower is the demand for access to the respective infrastructures and the lower is the supply of railway and airline services.

The precise nature of the charges for getting access to infrastructure will not be further elaborated on here. It is, however, important to emphasise that the situation which is considered in this chapter is characterised by a level of fees for infrastructure access which leads to a state of affairs where all demand cannot be accommodated. With this in the back of the mind, the current principles for handling scarcity and congestion in railways is described in section 4.1 and in section 4.2 for airports.

## 4.1 Railways

The time-tabling process in the railway sector is initiated by operators submitting their requests for the upcoming period. In this, each train is described in terms of point of departure and final destination plus an enumeration of in-between stops. The train is characterised by technical qualities (speed etc.) and weight which gives it a certain pattern of acceleration, free speed, deceleration and stopping. The calculation of these patterns are today computerised and generic patterns for each type of train (a passenger service between stations A and F, stopping at B, C, D and E; a freight train of certain weight and length between F and G etc.) are generated. For each train, the time it takes to use a block if it accelerates, brakes or travels at free speed is thus created and stored for subsequent use.

In addition, operators specify the preferred departure time for each service. The combination of demand and technical specifications is sufficient to describe the whole movement of the train between departure and arrival stations.

Since the going schedule is being operated and has manifested itself to be functional it is typically used as a point of departure for coordinating demand. The process also makes use of rules-of-thumb for prioritising different classes of services; high-speed

passenger trains may be given priority over standard passenger trains, which may be ranked higher than high priority freight services etc. Furthermore, formal CBA is used in Britain's medium term planning of capacity use when dealing with controversial open access applications for paths (Johnson & Nash 2008).[4]

In reality, priority changes in one part of the network may have consequences in other parts, making it impossible to base the trade-off on a fixed principle. At the end of the day, the skill of experienced staff is therefore at the core of the prioritisation process.

For reasons to be further described in section 5.1, there are no formal optimisation instruments available to establish railway timetables. When changes have to be made during the preparation process, and although software is available for creating graphical time tables on the screen, the basic analytical means is paper, pencil and ruler as a technique for understanding the consequences of gradual alterations of drafts. In addition, the sheer size of the problem makes it impossible to test more than a few alternative solutions.

Directive 2001/14/EC provides the possibility to sign framework agreements between an operator and the infrastructure provider. This type of agreement makes it possible for the operator to acquire an option to have a (degree of) right-to-capacity for a certain number of years in the future. In this way, a degree of stability or at least reduction of uncertainty is granted an operator investing heavily in rolling stock with low opportunity value. The downside of the framework agreement is that it reduces the flexibility in future track allocation processes, thus providing a challenge to efficiency. The reduced flexibility may, however, also make the prioritisation slightly easier since it becomes necessary to accept some of the demand from some operator(s) without confronting it with the wishes of others.

A mandatory component of the same Directive is for member states to implement a performance regime. This is a system for penalties (and possibly bonuses) on trains which deviate from the patterns they are allocated in the timetable. The purpose is to incentivise both the infrastructure manager and the operators to undertake their

[4] More information about how the process is handled in France is given by Perennes (2012) and in Sweden by Jansson & Lang (2012).

respective activities in a way which does not affect the possibility for others to run their trains according to the time table.

One way to moderate the vulnerability of a time table is to reduce the number of trains which are allocated slots. This could be implemented on a voluntary basis in so far as operators that are aware that capacity is strained may abstain from asking for one or more leg(s) of a train. More probably, it may be necessary for the infrastructure manager to establish an upper limit for the number of departures on a line. In this, the fact that knock-on effects from primary disturbances grow with the number of departures is an important datum. It may therefore be preferable to establish schedules in which capacity is not fully utilised in order to ascertain acceptable quality. The stricter is the upper limit set, the scarcer is capacity.

Gibson et al (2002) also reports that the UK has introduced a capacity charge. This part of the overall charging scheme is supposed to reflect the (expected) marginal congestion costs of track access. Since it differs with respect to where and when a service is operated, it provides motives for reducing demand for these slots, potentially having consequences also for the degree of scarcity. Also Sweden charges more for the use of more congested parts of the network[5], although the level of fees is still very low, i.e. may be ineffective as a means for affecting demand.

When it comes to the priorities once trains have been delayed, it is reason to consider the possibility that a service which was not prioritised in the scheduling phase would be more hurt by being delayed that another service which was given higher priority. The reason may, for instance, be that a freight train is flexible ex ante but once a slot has been allocated the production process at the final destination is scheduled to fit the arrival time. A sea transport may, for instance, be scheduled within a tight time window and a delay may mean that the boat has to leave with considerable consequences for the chain of deliveries and subsequent production. This feature has to be accounted for when priorities have to be settled for the implementation phase.

---

[5] http://www.trafikverket.se/Om-Trafikverket/Andra-sprak/English-Engelska/Railway-and-Road/Network-Statement1/Network-Statement-2013/

A further aspect of the treatment of trains which are late is that the only way to handle the situation is to *delay* a train relative to its scheduled slot. In contrast, the scheduling problem is handled by way of moving slots both forwards and later in time in order to identify a value maximising solution. This feature makes the congestion problem analytically simpler to solve since the number of alternative solution is drastically reduced compared to when trains are scheduled.

Törnquist (2006) discusses several aspects on the delay management problem. There are today an increasing number of mechanisms to make adjustments as delays occur, and indeed also to keep train drivers updated on the adjustments made on a real time basis. The basic principles used when making these adjustments still seem to be crude, one example being to always give priority to trains which are not delayed.

### 4.2 Airports

Council Regulation (EEC) No 95/93 established the principles for handling demand when charges don't clear the market for airport services. The procedures date back to a system created by IATA in 1947 which over the years has been gradually updated. The system is administrated by airport authorities, but it is the airlines themselves that have the last word about the allocation of available slots.

The rules state that arrival and departure slots shall be allocated by a coordinator for an airport when demand exceeds available supply. The coordinator is instructed to handle the task in a neutral, impartial and transparent way and is supposed to participate at a bi-annual international time-tabling conference. Except for one representative of each airport, spokespersons for all airlines take part in the meeting which may include well over 1000 delegates from more than 250 airlines.

As a prologue to this meeting, airlines submit their (confidential) demand for slots to the coordinator. Using the principles established by the Directive, these wishes are compiled and a draft proposal about the number and timing of slots allocated is sent back to each operator before the opening of the conference. A basic principle that steers this first allotment is that existing slots are grand-fathered, meaning that an airline that flies in a certain way during the current timetable is given priority to the same slot in the next time-table, it this is what it wants.

A slot has to be used at least 80 percent of the time during a time table period if the historical right to the slot during the equivalent next season is to be protected (the use-it-or-lose-it rule). A grandfathered slot is therefore not the property of its incumbent owner. The Directive also provides for the creation of a slot pool into which newly created slots through increases in hourly scheduling limits, slots returned either voluntarily or under the use-it-or-lose-it rule and slots otherwise unclaimed by anyone are placed. At least 50 percent of all slots in the pool shall be earmarked for entrants.

During the conference, the initial proposals are discussed between operators in order to establish a final solution to the shortage problem. This illustrates the strong position of the operators when the allocation is established. This in turn shapes the behaviour of the coordinator in the creation of the draft proposal. After that an allocation has been established there are also on-going bilateral negotiations between airlines that seek to establish mutually beneficial exchanges of slots.

# 5. Shortcomings of today's approaches

Against the background given in section 3 of criteria for maximising social welfare when a timetable is constructed, this section sets out to define the limitations of the current approach to capacity allocation. To this end, and using a railway example, section 5.1 describes two generic features of the capacity allocation problem. Section 5.2 and 5.3 make use of this in order to elaborate on the shortcomings of today's approaches in the rail and air industries, respectively.

## 5.1 The underlying problem

In order to provide an understanding of the challenges which have to be dealt with when a timetable is to be constructed, it is necessary to take a step back in order to provide a deeper understanding of the problem at hand. The construction of a train schedule is used as an example.

The objective in this process is to get the maximum social benefits out of existing infrastructure. The qualification that benefits should be considered from a social perspective emanates from the text in the Directive and – from a more profound perspective – from the fact that Europe's infrastructure is supported by substantial public sector funds.[6] It is therefore reasonable to require optimisation to consider more than financial flows.

A first challenge when addressing this task is to handle the optimisation problem, which concerns the mathematical aspects inherent in the problem. Formally, the infrastructure manager (IM) makes tracks available for $i=1,\ldots,I$ independent train operators. It is for the time being assumed that the value of each path $r$ to each operator $i$ is public knowledge, and so is also the value of alternative ways to run the service; $v_i(x^r)$ represents the value to operator $i$ as a function of the path.

The IM seeks to establish the value-maximising solution to (1), i.e., to find the highest aggregate value over all operators' values of all paths. The first constraint means that

[6] Directive 2001/14/EC provides the platform for the current Europe-wide policy vis-à-vis its railway industry. It sets out minimum requirements for member states, in this particular case concerning maximum level of charges.

at any time *t* there can be at most one train *r* using one and the same block *s*. With *X* representing all feasible paths, the second constraint indicates that the paths must be technically feasible to operate. Due concern must therefore be given (a) to physical limitations on rolling stock (acceleration capability, max. speed, etc.) and tracks (line-specific speed- and speed-/load restrictions) as well as (b) to established safety restrictions. In other words, the schedule must be possible to operate, should no one else ask for access.

$$\text{Max } B = \sum_i \quad \sum_r v_i(x^r) \qquad (1)$$
$$\text{S.t. } \sum x_{s,t}^r \leq 1 \quad , t$$
$$x^r \in X \; for \; all \; r$$

There are two aspects which contribute to the problems with establishing a solution to the optimisation problem. The first is its immense size and complexity. For each block, a train may be accelerating, going at maximum speed of be decelerating. The number of blocks and stations between origin and destination may be large, creating a huge number of alternatives when the train is alone on the line. Introducing one meeting train means that one of the two must be given priority when asking for access to the same slot. Each alternative solution generates a new set of alternatives subsequent to the conflict which has been solved. The complexity grows very fast when the number of (conflicting) trains increases.

The second complication is related to the binary restriction on the objective function which means that there may be zero or one train on each block at a time. The binary restriction makes the optimisation problem unfit to handle with traditional optimisation techniques. The reason that the railway sector does not use formal algorithms to establish timetables is thus that there today are *no optimisation techniques available for establishing solutions to the problem* which squeezes the most benefits out of the existing capacity. The problem is said to be NP complete, NP being short for non-deterministic polynomial time; cf. for instance Rothkopf et al (1998).

While there are no techniques available for establishing global optima, it is feasible to reduce the size of the problem in different ways. Brännlund et al. (1998) use so-called

dual optimisation or Lagrangian relaxation to handle the challenge for a part of Sweden's railway network. This approach establishes a solution which lies within a percentage interval around the underlying optimum. Using another approach, Borndörfer et al (2006) is further example of how the problem can be dealt with. Subsequent research has further enhanced the scope for establishing solutions of adequate quality also to complex network problems, "adequate" referring to equilibria for which the duality gap is not large.

The above description is based on the assumption that the IM is aware of the benefit of allocating path *r* to operator *i*, i.e. of $v_i(x^r)$. As an illustration in a very simple situation, this is necessary to strike the balance between trains A and B in Figure 1 on the one hand and C on the other. This is referred to as the incentive problem, i.e. the need to make operators reveal their value of track access.

In the traditional, vertically integrated business model for railways, where different divisions were parts of the same monopoly, it may have been reasonable to assume that such information could be acquired. In a (more or less) competitive situation, operators however have reason to argue that their services are highly valuable or that they would be severely hurt by not being allocated slots according to their demand specification. Even if it would be feasible to implement an optimisation algorithm, this could not be expected to deliver a value maximising solution if information about the value of single trains is not available or is of poor quality.

Not only the optimisation, but also the incentive problem is inherent in the time tabling problem and not related to market structure. Even in the traditional monopoly there may, thus, be internal conflicts between divisions competing for the same slots, making it difficult to induce division managers to submit truthful information about their value-of-access.

### 5.2 Railways

Confronting these analytical properties of the time tabling problem with the description in section 4 of the way in which schedules are derived today makes it obvious that present time tables cannot be guaranteed to be efficient. Within a given

market, services with more efficient marginal operations relative to competing services should expand. Today's procedures are however not designed in order to make operators to report about their value of access in combination with specifying demand. Rather, more or less explicit pecking orders, in conjunction with iterations between different solutions decide final priorities.

It is not unreasonable to assume that rules of thumb of this nature on average can provide a reasonable starting point for the process. But it is equally obvious that single, non-average departures may be given inappropriate priorities. Which is the proper prioritisation of a long distance passenger service relative to a crowded commuter train? A high-speed train on a return leg without so many travellers compared to a freight train with tight deadlines at its destination? A regional service with not so many travellers but of high value to local communities competing for priority with a cargo train taking paper rolls to a port? If these questions cannot be given well-founded answers, it is not reasonable to expect that operator and carrier expansion and contraction function in an efficient way. These question-marks obviously multiply when taking a system wide perspective of the challenges.

A further requirement is that the allocation procedure should be reasonably cheap to administer, including the oversight that is necessary to minimise the risk for collusion etc. Today's process is, however, extremely slow with almost three months from that requests have been submitted to a draft timetable being announced. Moreover, only one additional iteration is feasible before the final schedule is settled. As indicated above, the principles governing the final schedule are also intrinsically non-transparent.

With demand for railway services growing over time, the current administrative process does not provide any signals about the value of additional track capacity. The only information which emanates from the process is whether or not demand exceeds existing capacity.

If equity is a policy issue within the transport sector, it is typically the regional dimension of the trade-offs which have to be made which generates concerns. In most peripheral parts of railway networks, capacity is underutilised. But some long-distance

trains start in small communities and have their final destination in congested parts of the network. Considering the number of passengers in the respective trains, it is reasonable to expect that the long-distance train should be given priority if a conflict over capacity with a regional service would materialise. This conclusion may, however, not be obvious if distributional concerns provide a motive for giving priority to the regional service. The only way to handle this type of conflict is to ask policymakers to provide the planners with an explicit weight of regional versus long distance trains.

### 5.3 Airports

The generic model of section 5.1 can be used as a platform for understanding the challenges also for the allocation of airport slots. The fundamental difference is, of course, that the mathematical complexity of the problem is so much smaller. Although it is necessary to acquire not only a slot for departure but also a landing slot at another airport, the degrees of freedom in this coordination problem are vastly larger than when railway services are to be scheduled. It is therefore straightforward to handle the allocation problem of different airports as being independent, at least as a first approximation.

Although this eliminates most complexities of the optimisation problem, it does not change the nature of the incentive problem. The procedure to allocate slots between competing users which has developed over the years can be understood as the industry's approach to grapple with this problem. The approach can be seen in the light of the economic theory of clubs. The club is a voluntary group of individuals who derive mutual benefit from sharing production costs or getting access to a good characterized by excludable benefits. A *club good* is the sharing of an excludable (rivalrous) public good.

When club decisions can be represented as a cooperative action, the resulting outcome will be a Pareto optimum for the club members; members belong since they perceive a net benefit from membership. If not, they would exit. If the club decisions accounts for the well-being both of those within the club and those outside, then the outcome will be a Pareto optimum for the economy as a whole. Cf. further part IV in Cornes & Sandler (1996).

An airport is an example of a multiproduct club used by commercial and non-commercial passenger and freight operators. The question is then what the equilibrium outcome from this specific club could be expected to be. A first observation is that the possibility to start new clubs if supply is saturated at an existing facility is limited. Alternatives can at least be considered to be inferior to the preferred alternative which typically is an international hub (London/Heathrow, Amsterdam/Schiphol, etc.). It is therefore not reasonable to expect that efficiency can be achieved by way of saturating demand by expanding capacity elsewhere.

Considering the decision-making of a club at one specific airport in isolation, a further concern is whether or not it could be conceived of as a (fully) cooperative solution. If members are partitioned into two categories – large and small, or incumbent and entrant airlines – it is not obvious that they are handled as equals when decisions are to be made. A small or entrant firm could have developed a new type of service with substantial earning capacity, but can still choose to accept a slot with poor qualities as suggested by the other club members. The reason may then not be that the entrant conceives of it as overall beneficial but since the alternative – not to receive any slot at all – could be worse. Not least the feature of grandfathering provides incumbents with more influence in the committee meetings than they would have if all would start from scratch and try to find a solution to the benefit of them all. This is so since the behaviour of club members is driven by the rule in force – the grandfathering principle – if members can not arrive at a common conclusion.

It is therefore not clear that the cooperative solution could be thought of as a (fully) Pareto optimal allocation between operators that actually belong to the club. The club solution is even more socially problematic if the problem of becoming a member is considered; the position of many flag carriers at national hubs is strong compared to that of prospective entrants. Moreover, the haggling at international gatherings provides incumbents with a long history in the business with even more influence relative to an upstart. In short, there are few reasons to believe that the scheduling committees of major airports allocate slots in ways promote efficiency in the way discussed above.

From the perspective of inter-temporal efficiency, the club solution has an additional drawback. If a market would be set up for airport capacity, it is reasonable to believe that the price for airport slots would differ over the day, i.e. be higher at peak than at off peak. A high peak price means that at least part of the demand is induced to use airport capacity at off peak, meaning that demand is rescheduled over the day on a voluntary basis. The size of the peak premium that is charged does also work as an investment signal; at high enough price levels the operator would earn even more if airport capacity was expanded.[7] This is not so with the current administrative procedure. Rather than providing incentives to reallocate demand over the day, operators have reason to ask for additional capacity. (Demand for) over investment is therefore a complementary drawback of this approach.

7 This is at least so if the market is competitive since an overly high charge would otherwise induce entry into the business. Airports represent a less competitive activity so an uncontrolled monopoly supplier of the services may have incentives to under invest in order to hold up prices and earn supranormal profits.

# 6. Towards a more efficient allocation of slots

Current procedures to allocate railway and airport capacity have obvious shortcomings. This section addresses the challenge to build a more efficient slot allocation mechanism by starting with a general discussion about capacity allocation in the perspective of charging for the activities at large (6.1). Like before, the subsequent sections 6.2 and 6.3 then discuss potential means for adjusting the principles of today in railways and at airports, respectively. Section 6.4 then addresses the arguments for and against using operator profits as a proxy for social welfare.

## 6.1 Charging and capacity allocation

In any market, higher prices will reduce demand to a higher or lower extent. This means that higher ticket prices will reduce demand for air and railway travel. And since costs makes up the decisive determinant for prices, the charges for access to airport and railway infrastructure will affect demand for access and may have knock-on consequences for demand from passengers and freight customers.

Increasing the price for using railways or airports will therefore automatically solve the allocation problem since demand will fall back and all applications for slots can be granted. Uniform price increases may, however, deter from operations also in periods, or in parts of the network where capacity is abundant. This may provide one motive for restrictions on the level of charges in the railway industry. At airports, prices which clear the market may earn the owner monopoly profits, which could provide a complementary argument against full-fledged use of the price mechanism.

But it is also well established that the appropriate way for ascertaining efficient use of scarce resources is to levy charges which correspond to the (social) marginal cost for using them. This is, indeed also at the core of the current EU policy as established by Directive 2001/14. In this context, costs also include scarcity costs. Again, this means that the level of charges is not a par with marginal costs, since – if so – there would be no excess demand. Nash (2005) describes the current level and structure of charges in Europe.

The qualification that costs should be seen from a social perspective means that externalities such at the traffic's environmental impact, accident risks etc., should be made part of the charging scheme. It is therefore reasonable to consider the scarcity problem assuming that these concerns in one way or another have been dealt with. If not, these concerns should be addressed head on rather than by adjusting the way in which scarcity is dealt with. For the subsequent discussion, no attention is therefore given to the treatment of these wider, social aspects of the respective industries.

It is obviously no conflict between marginal cost pricing and a higher degree of price differentiation. A natural first step for handling scarcity problems would therefore be to increase the price during periods, and in parts of the (railway) network, where there is obvious scarcity while retaining prices in other parts of the network. Some examples of that this happens have already been given. But since there are many examples of non-satiated demand also in these countries, the differentiation is currently not sufficient to clear the market.

We can think of this approach as a system with posted prices where (inter alia) scarcity prices are published well before operators submit their demand for access for the upcoming time-tabling period. To simplify, we can think about a two-level structure with a higher price for peak, and a lower for off peak periods; the concept generalises up to any number of periods. Operators would submit their demand for departures, given this pricing scheme. The IM, be it an airport or railway authority, could then see whether or not excess demand has been eliminated. If not, the authority could raise the peak price for the subsequent time-table period; if the price is so high so that capacity is not fully used the price would have to be reduced.

This is the standard way for a Walrasian tâtonnement equilibrium process to work. Both theoretical modelling and experience could help to insure that the initially specified prices are not too far away from equilibrium. This is all the more important in view of the fact that at least railway time-tabling is a cumbersome process and that the first proposal for a time-table rarely is adjusted more than at the margin.

## 6.2 Improving capacity use in the railway sector

The case for price differentiation in networks with excess demand is obviously strong. The more uncertain is the value of different services, the more difficult it is to establish prices which clear the market without squeezing out demand on links where scarcity is not a major issue. An alternative approach could then be to ask operators to quote a price for the trains they want to operate. We have then moved over to an auction-type mechanism where a price is not posted in beforehand but rather established as an inherent part of the allocation process.

The scarcity cost in the railway network is defined to be the deviation from individually ideal timetables and its subsequent loss of revenue caused by insufficient availability of infrastructure capacity. The ideal timetable is thus the departure-arrival pattern that would prevail if everyone could run their services in the way they want, without obstacles in the form of other operators' services. Scarcity is manifested either in that trains have to depart at other times than they would prefer or in that they are completely closed off from operation.

Based on these considerations several parallel and sequential research projects have developed auctioning mechanisms for allocating track capacity.[8] Core features of this process include the following aspects:

Operators interested in running train services are asked to articulate their demand with some more detail than they do with present techniques. In particular, each operator specifies their demand not as a single point in time but rather as a time interval. A long distance train may, for instance, require two departure slots for the morning peak between 6.30 and 8.00, and there should be at least 45 minutes in-between the departures, but with these restrictions there is a degree of flexibility which could be accepted. An operator of commuter trains that leave every 15 minutes does not care if the first train departs on the hour, one minute past, two minutes past etc., only that the whole set of departures leave with a 15 minutes gap in-between. And a freight service could be loaded and ready to leave at 4 pm but given the needs at the location, it may

[8] Brewer & Plott (1996), Nilsson 1999 and Isacsson & Nilsson 2002. A comprehensive description of the process is given in Nilsson (2002) and the optimisation technique that has been used is presented in Brännlund et al. (1998).

be feasible to accept any departure between 4 and 8 pm. The more flexibility built in with the original specification of demand, the more latitude is given to the subsequent process of adjusting demand to existing capacity.

In addition, the operators are asked to specify a value or a bid function over these departure times. The ideal departure time from station A is perhaps 07:00 but it would be feasible to leave anytime between 06:40 and 07:20. The ideal departure may be worth 100, but this value could taper off to zero at the outer boundaries of the feasible schedule; this is the continuously falling distribution in figure 2. Alternatively, their valuation may be identical irrespective which departure slot they are given within the time interval, which is represented by the rectangular distribution.

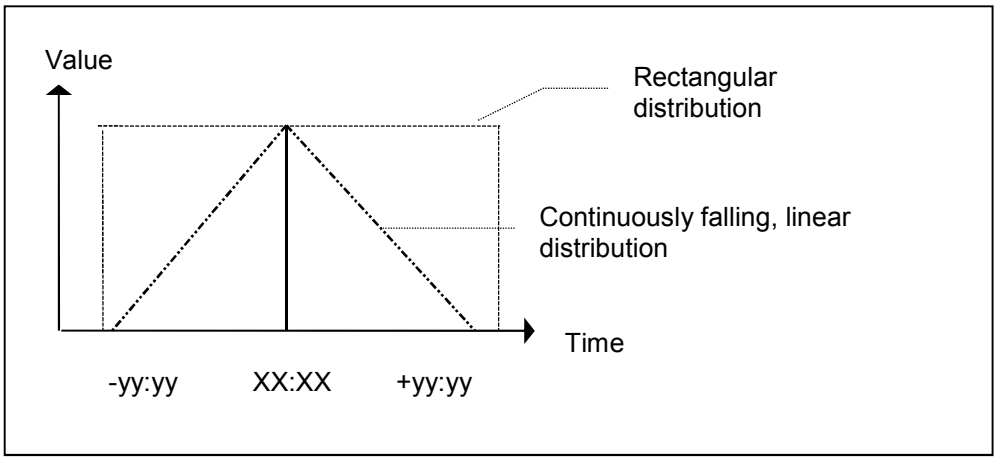


**Figure 2:** Examples of how value-of-access may vary with departure time.

Demand data of this nature is fed into an optimisation programme that establishes the value-maximising solution to all wishes. Operators are informed about what departure time that they are allocated – if any – and what they would have to pay in order to run their train in this way; they may then revise their demand specification. The process goes on as long as anyone would want to make changes of their demand scheme.

This process provides a solution to the capacity allocation problem and at the same time generates a set of prices. This is the scarcity charge that would make up a

separate component of a comprehensive pricing scheme. We can note a few qualities of this way to establish a price for capacity use:

- The charge is made operational by specifying the payment for each specific origin-destination pair.
- The level of the charge depends on the level of demand relative to available capacity in each specific part of the network. A certain A-to-B train may have to pay different prices during different parts of the day. Another train between B and C which in all other respects is identical to the first (type of rolling stock, length of trip etc.) may have to pay a different charge.
- A railway network, and the links between trains making use of it, provides an intricate cobweb of interactions from a time-tabling perspective. Even if only very few operators would be using tracks it would be very difficult to coordinate demand in order to fool the system.

It is important to emphasize that this description bypassed the major obstacle to full implementation of this approach, namely the fact that an optimisation mechanism may not yet be available. It does, however, provide a benchmark for the type of information which is vital for establishing an efficient use of the network.

### 6.3 Improving capacity use at airports

Research into the economic theory of auctions has virtually exploded over the last few decades, and the references given above represent work within this tradition. There is still reason to go back here to the seminal work relating to slot allocation at airports. In 1979 the US Civil Aeronautics Board and the Federal Aviation Administration commissioned a study to find appropriate ways to deal with shortages in airport capacity. The subsequent report was not made generally available until 10 years later; Grether et al. (1989) was suitably published in a series referred to as Underground Classics in Economics.

The suggested mechanism comprised the following features. The (primary) market for slots was organised as *a sealed-bid, one price auction*. Each potential buyer submits a bid for each slot in demand, indicating the maximum price the buyer is committed to

pay. Bids are then arrayed from highest to lowest. If x units are to be auctioned, then the highest x bids are accepted. The price paid by each of the winning bidders is the value of the lowest accepted (or highest rejected) bid.

The research established that these rules provide each buyer with incentives to bid (close to) the maximum that he/she is willing to pay. This is of course directly related to the profits the flight will generate. As a result, the economic circumstances are reflected immediately and accurately in the market; slots are allocated to those airlines which value them highest. For carriers it means that the profits from their most profitable flights are protected – they will not be dissipated for slot acquisitions; the highest bids do not determine price.

The proposal suggested that the sealed-bid auction would be supplemented by an *aftermarket*. The reason was that the sealed-bid auctions could be applied only to one airport at a time. The aftermarket would then handle the coordination between airports.

There may be a risk that services to *small communities* will be terminated if slots are allocated by a market process. Markets can, however, be organised in ways which will prevent this from happening. One way to do so is to establish a restricted market for small communities. Out of a total number of slots, a certain number would then be restricted for this use. Only buyers with special status could participate in the bidding for these slots. The special status could alternatively be based upon the origin or destination of the flight, passenger classification of aircrafts (commuter vs. general aviation etc.) or the size of the aircraft.

Certain types of aircraft use more “capacity” than do other aircraft. An increase in the share of “heavies” would then result in a loss of airport capacity. This could be handled by a flexible way to deal with *slot definitions*. Operations that have characteristics which place disproportional demand on capacity require more slots than other operations and it would become more costly to bid for them. The auction could handle this by making it necessary to bid for two slots for large aircraft or to make the precise definition part of the mechanism.

*Disposition of funds*. Funds generated by the sale of slots should be used to defray the cost of removing the binding airport capacity constraints. Many possibilities exist, including the establishment of satellite airports, but almost all of them require funding. The sale of slots provides a natural and economically efficient way of recovering the costs.

*Antimonopoly policies*: It is difficult to see how a carrier successfully could utilize an auction process to monopolise an airport. Even collusion is difficult since the auction rules could be designed so that neither winners nor bids are announced. The substantial number of airlines that normally operate at an airport would per se jeopardize the success of any collusion attempt. A possible aftermarket for trading could also be organised so that neither buyer not seller of slots need to know the identity of each other. But monopoly or predatory behaviour is especially difficult since the act of driving up slot prices to prevent competition necessarily uses up some or all the presumed monopoly profits. Furthermore, with the above structure, the funds would be destined for capacity expansion which would further undermine any monopolistic tendencies.

Monopolistic tendencies, if they occur, could also be managed by the use of complementary rules. A monopoly would be effective only if it could withhold supply. In the case of airports this would mean that large proportions of slots go unused or that they are used for operations which do not involve many passengers. Revenues from several of the operations would then not cover the price paid for slots. The existing use-it-or-lose-it rule could make attempts to pre-empt competitors very costly.

The market mechanism was never implemented, the main reason being – as described in the preface to the Grether et al report – the fierce resistance from the airline industry. From April 1986, a different type of market based mechanism, known as the "buy-sell-rule", has however been in use at four major US airports (Kennedy, LaGuardia, O'Hare and National). The baseline way to allocate capacity is still via committee procedures based on grandfathered rights. The new rule lays down that, in addition, any person is authorised to purchase, sell, trade or lease slots. It is thus

feasible not only to buy and sell the right to a particular slot permanently but also to lease them on a temporary basis.

### 6.4 Profit or welfare maximisation?

It has so far been assumed that the use of bids by profit maximising railway or airline operators provides a reasonable proxy for social welfare. There is finally reason to acknowledge the limits of this claim.

The basic argument in favour of using this approach is that any operator of passenger services is anxious to provide services which the customers appreciate. This is indeed the most basic way to retain existing and lure new travellers to using the service. It also includes quality aspects such as the service being on time, being clean and tidy etc.

The previous review has also assumed that any environmental consequences of a flight or a train are appropriately priced. A bid from a profit maximizer would then not be compromised by market failure of this type. But this assumption may also have to account for that prices in other modes of transport don't fully include all external effects of these services. The obvious recommendation is, of course, to change the pricing policy in these other sectors. If this does not happen it provides a motive for making second best offsetting adjustments in the pricing of capacity.

Nilsson (1992) discusses some second best aspects of infrastructure pricing. One obvious recommendation in this is to reduce (increase) charges in sector A relative to their first best level if mode B is priced below (above) its first best level and if this – for some reason – may not be adapted to the first best level. To balance the effects on demand of prices below (above) the first best level, investment in both modes should be smaller (larger) than in a first best world.

In the same way, it is reasonable to argue that monopoly behaviour – if it is believed to exist – is better handled head on, i.e. by breaking up a monopoly, than by complicating the treatment of bids for scarce capacity. The use of complementary rules such as use-it-or-lose-it will also make it costly for a bidder who considers the possibility of pre-empting entry by buying slots which it does not intend to use. This,

and possibly also other rules relating to the allocation mechanism, provides a complementary reason for believing that the risk for excessive market control provides an argument against using bids as an input. Borenstein (1988) however provides a critical argument against relying on bids for allocating scarce capacity.

In practice, it is reasonable to expect that the value of a slot to an operator is strongly correlated with the number of passengers in a passenger train or by the value of the cargo in a freight train. A complication may appear if some (commuter) services with many passengers are not operated on a commercial basis but paid for by some public sector representative. Commercial operators could then complain over having to compete with a service provider with deep (public sector) pockets while the tax payers may find it un-fair and costly to have to bid against the commercial firm. It is, however, not obvious that this type of concerns should be allowed to affect the allocation of slots.

These and other types of fears are frequently raised against using bidding mechanisms at large. But irrespective of this, it is still feasible to use the bids as a point of departure for establishing the optimising solution. By using pre-announced adjustment weights, some bidders or bids on some types of services could be given an artificially higher value, increasing the chance of being allocated a valuable slot. This was used as part of the US auctions for radio wave frequencies in the 1990ties (McAffee and McMillan 1996) where for instance bids from radio stations operated by minority groups were given preferential treatment. This points to a technique which could be used irrespective of which bid corrections would be seen as important by a policymaker.

## 7. Conclusions

Excess demand provides a signal of inappropriate pricing, i.e. that prices which clear the market are not in place. This gives a strong argument for higher, or at least more differentiated prices. Differentiating prices according to place and time is also cheap to administer and it has a major potential for improving the use of existing resources in several dimensions; demand may voluntarily be moved in time and room; operators are given reason to add cars to a train rather than running more trains, and so on.

As for airports, resistance from the industry has been decisive in order to block the use of the pricing mechanism to clear the market. Slots for take-off and landing are the main asset for airlines to earn money, so the willingness to relinquish their grandfather's rights is fierce. The consequence is that entry to airports which are highly used is difficult for entrants, providing one explanation for the growth of secondary, cheaper airports with a lower degree of demand relative to capacity and often lower charges.

There are examples of differentiation of charges relative to scarcity in the railway industry. It is, however, obvious that the level and degree of differentiation is not sufficient to close the gap between demand and supply. Except for increasing prices, there is also a discussion in both industries to use auctions to equilibrate supply and demand. In that way, the degree of differentiation could be established directly by supply and demand.

As the situation stands right now, with poor optimisation devices available for establishing equilibria, it is not feasible to create nation-wide solutions to optimise railway timetables. The reason is the immense mathematical complexity of the problem. Piece-wise optimisation for parts of the network is, however, clearly possible. This points to the direction to go irrespective of if iterative bidding is the preferred solution to establish relative priorities or if ex ante relative weights would be used. The cumbersome methods of today, which for instance only facilitate very limited trial-and-error before a time table is established, do by themselves provide a strong argument in favour of further development in this direction.

The substantial complexity of the market, the heterogeneity of demand and the blend of time and railway lines with both inappropriate and excess capacity, indicates that the potential loss of value is high in the railway industry. Not least in systems working close to capacity more elaborate techniques for making use of existing assets may make it possible to postpone investment projects. In addition, it would provide a means for utilising the existing capacity better than at the outset.

## References


Borenstein, S. (1988): ‘‘On the Efficiency of Competitive Markets for Operating Licenses.’’ Quarterly Journal of Economics, May.

Brewer, P.J. & Plott, C.R. (1996). A Binary Conflict Ascending Price (BICAP) Mechanism for the Decentralized Allocation of the Right to Use Railroad Tracks. International Journal of Industrial Organization, Vol. 14, No. 6, pp 857–886.

Brännlund, U., Lindberg, P.O., Nilsson, J-E & Nöu, A. (1998). Railway Timetabling Using Lagrangian relaxation. Transportation Science, Vol. 32, No. 4, November.

Borndörfer, R., M. Grötschel, S. Lukac, K. Mitusch, T. Schlechte, S. Schultz and A. Tanner (2006). An Auction Approach to Railway Slot Allocation. Competition and Regulation in Network Industries, 7 (2), 163-197.

Cornes, R & Sandler, T. (1998). The Theory of Externalities, Public Goods and Club Goods. Cambridge University Press.

COUNCIL REGULATION (EEC) No 95/93 on common rules for the allocation of slots at Community airports

Directive 2001/14/EC of the European Parliament and of the Council of 26 February 2001 on the allocation of railway infrastructure capacity and the levying of charges for the use of railway infrastructure and safety certification

Gibson, S., G. Cooper and B. Ball (2002). The Evolution of Capacity Charges on the UK Rail Network. Journal of Transport Economics and Policy, Vol. 36, Part 2, May, pp. 341-354.

Grether, D., Isaac, M & Plott, C. (1989). The Allocation of Scarce Resources. Experimental Economics and the Problem of Allocating Airport Slots. Underground Classics in Economics, Westview Press.

Isacsson, G & Nilsson, J-E. (2002). Re-regulation of Previous State Monopolies. An Experimental Comparison of Allocation Mechanisms in the Railway Industry. Working Paper (2000).

Jansson, K., H. Lang (2012). Rail Infrastructure Charging: EU directive, Swedish Concerns and Theory. Research in Transportation Economics, August.

Johnson, D., C. Nash (2008). Charging for Scarce Rail Capacity in Britain: A Case Study (2008), Review of Network Economics, Vol. 7, Issue 1.

McAffee, R.P., McMillan, J., 1996. Analyzing the airwaves auction. Journal of Economic Perspectives 10(1), 159–175.

Nash, C. (2005). Rail Infrastructure Charges in Europe. Journal of Transport Economics and Policy, Vol. 39 (3), September, pp. 259-278

Nilsson, J-E. (1992). Second Best Problems in Railroad Infrastructure Pricing and Investment**,** Journal of Transport Economics and Policy, September 1992.

“ (1999). Experimental Evidence on the Use of Priority Auctioning in the Railway Industry. International Journal of Industrial Organisation no. 17, pp 1139–1162.

“ (2002). Towards a Welfare Enhancing Process to Manage Railway Infrastructure Access. Transportation Research, Part A.

Perennes, P. (2012). Can the “Invisible Hand” Draw the Railroad Timetable? Paper presented at…

Rothkopf, M., A. Pekec and R. Harstad (1998). Computationally Manageable Combinatorial Auctions, Management Science, 44 (8), 1131-1147.

Törnquist, J. (2006). Railway Traffic Disturbance Management. Doctoral Dissertaion Series No. 2006:03, School of Engineering, Blekinge Institute of Technology