Forecasting robustness of timetables

W. Barter

First Class Partnerships, UK

Abstract

Ultimately, the timetable is proof of the capacity of a rail network. If a train can be pathed, then capacity exists. However, a crucial variable is the extent of spare capacity or “slack”, intended to allow recovery from the minor disruption that is inevitable in real life. Fundamentally, the ability to absorb some disruption trades off against full exploitation of nominal capacity.

The number of trains that can be built into a timetable also depends on many factors apart from the technical characteristics of the infrastructure and trains, such as the mix of trains and their stopping patterns. Usable capacity therefore varies depending upon how it is used to reflect commercial and political requirements. In the UK, these requirements are increasingly embodied in the Strategic Rail Authority’s Route Utilisation Strategies, which aims to prioritise allocation of route capacity whilst preserving robustness. Whilst these are clearly desirable objectives, the risks include the impact on fringe services and weakening of the link between the market and the train service.

To support such a policy and other strategic decisions, often years in advance of finalisation of a specific timetable, a means of forecasting capacity is essential. Any valid measure of capacity must reflect not just the hard capabilities of the infrastructure and trains, but also the business and political context and the required robustness of the eventual train service. Probably no single universal statistic can do this.

Clearly, anything that can reduce uncertainty in forecasting capacity and robustness will allow a better balance to be struck between robustness and over-provision or under-use.

Keywords: operations, timetables, scheduling, robustness, simulation.

1 Introduction

The years since privatisation of the UK rail network have seen significant increases in both numbers of trains run and numbers of passengers. For instance,
in May 1995 the number of trains on a typical weekday was 16,982, whereas the equivalent figure for May 2003 was 18,725, an increase of 10%. Meanwhile the number of passenger journeys in 2003 reached one billion for the first time since 1961, whilst annual passenger kilometres stand at a higher level than at any time during the whole period of nationalisation.

There are many reasons for this. Background economic growth is certainly part of the explanation, especially in the London commuting area. However, capturing growth in demand to rail rather than alternative modes does not follow as a matter of course, and marketing activities of the privatised rail companies must be given credit. Then, statutory regulation of fares was introduced with privatisation, and increases in regulated fares have been limited to Retail Price Index minus 1%. At the same time tax on petrol has been deliberately increased which, together with the level of world oil prices, has resulted in the cost of rail transport falling relative to that of the private car. And the act of privatisation, with all the associated publicity, adverse or otherwise, has itself raised the profile of rail as a transport mode.

Of course, the increase in the number of trains has itself increased demand. Off-peak services are now generally close to peak levels, as stimulating off-peak demand is the principal strategy of operators in commuter-dominated areas, where demand is effectively set by economic activity and season ticket rates are regulated. On the Cross-Country network, a major investment in new rolling stock has been accompanied by a strategy of running shorter trains at more frequent intervals. It is this that largely underlies the observation that passenger-km have grown less than train-km, especially for long-distance operators [1].

This growth in demand comes after a period of decline in network capacity. Apart from simple economy measures, many of British Rail’s investment schemes involved capacity reduction to minimise capital expenditure. Lines at the fringe of the London commuter network, when electrified, were singled to save costs of electrification equipment or gauge clearance for modern vehicles. The number of trains on the East Coast Main Line, electrified in 1991, is in places limited by the spacing of feeder stations, and reliability in high winds is jeopardised by the spacing of support structures.

The result is a network experiencing high utilisation, on which train service punctuality is a matter for concern. Amongst the responses of the rail operating companies is an increasing focus on the “robustness” of the timetable.

2 Timetable robustness

The robustness of a timetable is the extent to which it maintains its form in the face of adverse and unplanned incidents.

Given that the running times and train separations on which the timetable is based are correct, and that no mistake has been made in compiling the timetable, if all trains start their journeys on time and no incidents are experienced on route, then all trains will run on time and end their journeys on time. This, of course, very rarely happens.
If trains are timed precisely to minimum running times and dwell times, any incident on route inevitably leads to delay. Once a train runs late, it may delay others, particularly at junctions. Each train delayed in consequence may then delay further trains, so that delays escalate in a chain reaction.

The susceptibility of the timetable to this escalation of delays is essentially a function of:

- the number of trains in the timetable, determining the probability that one late-running train will delay another at a point of conflict;
- the potential for escalating delays presented by the track and signalling layouts, for instance, in the form of conflicting routes at junctions;
- associations between trains deriving from crew and vehicle workings, and from commercially-dictated connections.

Recent studies by Network Rail [2] have identified clear relationships between intensity of capacity utilisation and escalation of delays, typically indicating a critical utilisation level above which escalation becomes severe. Such relationships, whilst similar in shape, are specific to each route, reflecting planning issues such as the extent of allowances and slack in the train timings, and the complexity of the infrastructure (in terms of the prevalence of points of conflict in the track layout), presenting opportunities for one train to delay another.

Spare capacity might deliberately be left to minimise escalation of delays, whilst slack in running times will allow some recovery from delays. In the former case a trade-off exists between the benefit of running additional trains and the cost of the additional delays. The latter factor must be judged in the light of extension to advertised journey times, which will deter patronage. The business need is for a means of quantifying both capacity and usage of capacity so as to identify criteria for appropriate levels of utilisation.

2.1 What is capacity?

In essence, the capacity of a railway is its maximum possible usage. Utilisation is then the ratio of Usage to Capacity.

In reviewing different methods for measurement of capacity, two different approaches become apparent:

- Capacity is identified in terms of potential usage of the railway, for instance, by a statistic such as “trains per hour”. This can then be compared directly with actual usage measured in the same terms;
- Capacity as such is not identified, but by definition is “60 minutes per hour” for each component of the network. Usage is measured in terms of the time for which a train or group of trains occupies the network or a component of the network.

In the former approach, assessment of usage is straightforward, but identifying a valid indicator of capacity is fraught with difficulty. In the latter approach, the question of measuring capacity is effectively bypassed, but rather more effort has to be directed at measuring usage in order to convert “trains” to “time”.
2.1.1 Approaches to capacity measurement

The line headway sets a cap on the number of trains that can run, so at best the number of trains in the eventual timetable cannot exceed the sum of the possible trains on each of the lines in the network. But trying to mix trains of differing speed capabilities, or stopping and through trains, on the same line results in loss of capacity.

The conventional way of quantifying Utilisation in this case is to define a “Standard Path” as being that occupied by the dominant type of train. The Capacity is then regarded as the number of paths available for such trains. Other types of train are then said to use a number of standard paths, and the Usage is the sum of the standard paths for all trains run.

However, this analysis may lead to a pessimistic view of System Capacity. By running slower trains consecutively, the total loss of paths for fast trains is reduced, as some of the lost paths are common to more than one such train. This is known as “flighting”, and is exemplified by Eurostar services which operate as two pairs of trains in each hour so as to conserve capacity both on the complex network of routes South of London and in the Channel Tunnel itself. It is this scope for mitigation that makes estimation of the real impact of the traffic mix difficult.

An alternative approach, still based on planning headways, but avoiding the flaws in the Standard Path assessment, can be termed “Timetable Compression”. In effect, over a section of route and for a nominated time-period, all trains are “retimed” as early as possible, until each is running at a minimum separation from another at some point on its journey. Inspection of the train graph for, say, an hour of service would then show a cluster of trains at the start of the hour, with all unused capacity concentrated at the end of the hour. A quantitative indication of the utilisation is obtained by comparing the time-span of the compressed timetable with the original time-span. This reflects the existing mix of traffic and order of trains – by implication further trains could be run in proportion to the identified spare capacity assuming the same mix.

It should be noted that the utilisation of capacity, measured by either the “Standard Path” or “Compression” methods, is sensitive to the length of route that is selected for study. If only a short section is considered, loss of capacity due to the mix of speeds is less apparent. On the other hand, if a longer section is studied, loss of capacity will be exposed, but it becomes difficult to reflect the impact of mitigating measures such as regulation mid-route.

Both assessments, employed in isolation, have a further fundamental flaw. The examples above consider trains flowing through a “corridor”. However, at each end of that corridor there will probably be a station with a complex track layout, be it a terminus with a number of platforms and “throat” pointwork to connect them to the main line, or a through station fed by a number of branches which converge at a flat junction. The frequency with which the terminus can receive and despatch trains will derive from the number of platforms and the time required for alighting, servicing and boarding.

Thus, constraints arising at locations at each end of a corridor may mean that they do not have the ability to feed trains into a corridor at the same rate that they
can be passed along it. Consideration of the capacity of these “nodes” is an essential match for the analysis of the operation of the corridor.

2.1.2 Capacity of nodes
How capacity is actually used is important in considering a corridor. For a node it is vital, for instance, because the number of trains that can be run through a complex area depends very largely on how the timetable exploits potential for parallel moves. Moreover, it is frequently found that within a complex layout the service pattern leads to concentrations of trains at particular points within the layout. These points may then form the binding constraint on the number of trains that can be run.

Such concentrations can be identified simply by mapping the number of trains to the various “links” in the layout, where a “link” is a section of line between turnouts. Figure 1 shows a simple example.

![Figure 1: Trains per hour – usage of links in a track layout.](image)

Usage is thus easily assessed. However, the capacity of a link is not easily expressed in terms of Trains per Hour, as the time for which passage of one train denies the link to others will depend on the precise move being made. For instance, in Figure 1, a train arriving and entering the Bay platform will lock the critical link some distance from the station, so as to run in without restrictive aspects. Then, given the need to brake to a stand in the platform, the time taken to travel this distance and release the link will be relatively high. Conversely, a departing train will lock it just before departure, and release it very soon after departure.

A preferable solution, therefore, is for numbers of trains on each link to be factored by a link “lock time” – the interval for which its use by one train denies it to others – so as to derive a total locked time for the link.

Lock times might be derived from margins laid down in the Rules of the Plan. However, this is very much an approximation, especially where the link in question is some distance from a station or timing point, simply because the margins quoted are in the form of differences between arrival and departure times at a station.
A much more valid approach utilises simulation. In a simulation model, all track circuits in a complex would be included, together with data on signalling and on the traction units and trains. During the simulation run, times of locking and release of track circuits will be generated for each train, leading to a total locked time for each track circuit. This lock time will extend from the time at which a signalled route over a track circuit is set, well in advance of the passage of the train so as to model running on clear aspects, to release of that track circuit by the tail of the train. The difference between those times will accurately reflect variations in train speed as modelled in the simulation.

Whilst simulation studies can be time-consuming, this approach bypasses one of the major tasks of a simulation, that of preparing a conflict-free timetable, entering it to the model, and checking its correct operation before running simulation trials. Whilst a notional timetable will need to be entered, this can simply be an hour’s-worth of trains, sufficiently separated to ensure that they do not conflict with each other. The order in which they run is irrelevant, as the total locked times will be the same whatever order the trains run in, and whatever their specific times might be in due course.

The locked times established in this way will be a bare minimum representation of usage, with nothing that can be called “slack”. The figures are thus not open to challenge on the basis of including “operator’s padding”, but this leaves the problem that common sense suggests that locked times of 60 minutes per hour could not be achieved in practice. For purposes of establishing a reasonable criterion of feasible usage, benchmark figures have been obtained by simulation of areas of the national network that are regarded as, for all practical purposes, “full”. From these trials, 40 minutes per hour seems not to be exceeded except at odd points on the most intensively-used networks, with 20 minutes per hour a more normal figure.

Whilst this may seem low, it should be remembered that having a free track circuit is of no value unless other track circuits around it are free so as to allow a usable route to be set. One track circuit will often prove to be the binding constraint in a layout, and identifying this will suggest remedial measures.

### 2.2 Route Utilisation Strategies

In the UK, whilst growth in demand, and service planning aiming to stimulate demand by increasing frequency, lie at the root of much of the observed increase in usage of the network, some other causes are less reputable. Some new services are said to have been introduced to sterilise capacity that might otherwise be used by a competitor. Others have resulted from attempts to benefit from the revenue allocation system rather than to increase the overall revenue.

The Strategic Rail Authority has embarked upon a programme of Route Utilisation Strategies, intended to provide a basis for allocation of capacity between operators. This programme was prompted by the high utilisations observed after a period of effectively “free for all” allocation of capacity, and in particular doubts as to the reliability of the railway.

The Route Utilisation Strategies have adopted a measure termed the Capacity Utilisation Index (CUI). This derives effectively from a Compression approach,
and thus reflects the limitations of that methodology. However, the principal fault of that method, that nodes between corridors are not considered, is effectively overcome by identifying relationships between capacity utilisation and delays for specific routes, and by implication, for specific infrastructure configurations. It follows from this, of course, that changes to the infrastructure configuration will invalidate the identified relationships, and therefore that changes to robustness arising from such infrastructure changes cannot be forecast by this method.

It has been observed [2] that when utilisation, as measured in terms of the CUI, reaches 75%, delay “by reaction” becomes severe. Indeed, the overall benefit of the extra trains may be outweighed by their adverse impacts.

The critical “75%” figure for acceptable capacity utilisation at first sight seems to conflict with the findings from link analysis by simulation referred to earlier, which finds utilisation levels in the range 33% to 67% on layouts that are effectively incapable of accepting further trains. The explanation lies in the fact that the CUI is based on planning headways, which themselves contain an element of slack. So if, for instance, the planning headway is twice the prevailing signalled headway, utilisation of capacity is already limited to 50% of the theoretical level. 75% utilisation as measure by the CUI would be equivalent to 37.5% utilisation as measured by more refined analysis, and thus fully compatible with other measures. The message here is simply one of comparing apples with apples – decision criteria in terms of any given measure of capacity must be related to observations from actual operation made in the same terms.

2.3 Robust planning

Assessment of usage of capacity by means of Standard Paths or Compression reflects the impact of differing train speeds within a corridor, but neglects the ability of stations and junctions to pass trains between sections of route. Analysing usage within a complex track layout in turn neglects the impact of differing train speeds on the approach routes, and other inherent relationships dictating the precise pathing of trains.

In effect, it is not possible to predict with certainty how many trains might be run in the future on a given track layout until a timetable has been prepared. Even when that has been done, however, it remains unclear how reliably that timetable will operate.

The train plan comprises a complex set of dependencies, between trains on the network, sets of vehicles at termini, crews and vehicles, and crews and their reliefs to whom they will hand over their train at the end of a shift. Any of these dependencies is exposed to a risk of failure, which will jeopardise punctuality. Network Rail has estimated [1] that half of delays arise “by reaction”.

Apart from unplanned junction conflicts, if a set of vehicles arrives late at a terminus, it is likely that its next working will start late. The situation is exacerbated if the crew from the arriving train are to work a different next working, as this will also be delayed. At worst, the vehicles, driver and conductor will go three separate ways, so that one late arrival leads to three late departures. This situation might arise where every possible economy in
diagramming is being sought, at the expense, deliberate or otherwise, of punctuality.

The behaviour of the complex system that is a railway, in the face of “events” and dependencies, is beyond the scope of manual analysis. However, it is a natural subject for simulation, employing a computer model of the dependencies with input on the probability and magnitudes of disruption events.

2.3.1 Exploitation of simulation

Given this apparent suitability of simulation in railway contexts, it is in fact surprising how relatively limited actual applications have been. The principal applications of simulation have been in specific contexts that allow for data-hungry, lengthy and cross-disciplinary analysis, such as major infrastructure projects, franchise bids, and one-off special exercises such as those that support the Route Utilisation Strategies. In the context of timetabling, the cost and complexity of classic simulation projects, as well as limited access to appropriate data and skills in the fragmented industry, are deterrents [3].

It might also be argued that the need for predictive robustness of timetables, as perceived by a franchised operator, is not significant. Timetables are increasingly set by the Strategic Rail Authority. In any case year on year changes are not normally dramatic, unless prompted by a major project or franchise renewal. Moreover, robustness of the timetable is but one factor in overall service performance, and an operator may gain more from attention to issues such as rolling stock reliability and control responses.

Whilst valid, these arguments leave scope for exploitation of simulation in principle, so long as a system can be found that is not only valid, but also capable of being used within realistic timescales and with limited skill-sets and data. Contexts in which a train operator should find simulation of value are suggested to include:

- Routine timetable planning: even within similar levels of service there is a risk that changes in train order may have significant impacts on robustness;
- Resources programming: a major source of secondary delay may be associations between trains arising from crew and vehicle workings, and an operator may be faced with a choice between economy and robustness;
- Business planning around performance regimes: although the timetable to be operated, and many factors underlying service performance such as infrastructure failures, may be outside the control of a franchised operator, valid predictions of their impacts will support budgeting for performance regime payments.

Perhaps the greatest opportunity, however, lies with short-term services such as engineering blockades. Recently, the tendency has been to increase the use of blockades lasting several days or even weeks for major engineering work, as an alternative to a large number of short-term possessions. Such blockades maximise productivity, by reducing lost time spent on setting up worksites and restoring the railway to traffic, and by offering more flexibility within the worksites for planning tasks. However, they present a high risk to service performance, as the natural tendency is operate a service as close to normal as
possible on the restricted network, thus by definition increasing utilisation. Moreover, high-value residential and business traffic is more risk, compared with confining engineering activity to nights and weekends. There is much to be gained, therefore, by even a coarse quantitative analysis of robustness of the temporary service, to support decisions on level of service and associated resources.

In this context of changes to the network, CUI-based forecasts of robustness become invalid, and moreover will fail to identify specific problems within the limited network and short-term service.

Some trials have recently been conducted in Norway with modern train planning software, essentially the same system as used by Network Rail in the UK, that also supports analysis of the robustness of the timetable. Essentially, some changes to train timings will be randomly imposed, representing late starts or delays on route. Trains running out of their planned paths in consequence may then encounter conflicts at junctions that are avoided in the planned timetable. These conflicts are simply resolved within the software on the basis of preserving Rules of the Plan separations, so that the total delay may escalate, offset by allowances in the schedules.

The software in question [3] overcomes the major hurdles to exploitation of simulation by a franchised operator, effectively confining skills and data requirements to those of use of the system for train planning, and thus limiting the elapsed time required for analysis. Interfaces with resources scheduling systems will capture these associations between trains.

3 Conclusion

There is probably no single statistic to describe utilisation or capacity appropriately in all contexts. The issue is the number of trains that can run over a network of lines with conflicting routes at junctions, at the times when the public, politicians and SRA want them to, despite their different speed characteristics, and at an acceptable level of reliability in practice.

This Paper has discussed a number of basic measures of capacity and usage:

- “Trains per hour” - only applicable to individual lines or very simple systems, does not consider the traffic mix;
- Standard paths – neglects the capacity of Nodes, may be distorted by the order in which trains run;
- “Timetable Compression” – sensitive to train order, but neglects the capacity of Nodes;
- “Link Usage” – neglects corridor effects on route.

Any of these methods may exploit planning headways and margins as opposed to simulation-generated minimum figures for headways and link occupations. Either approach is valid so long as both capacity and utilisation statistics are derived in the same way, and benchmarked against observations in the same terms. For instance, usage assessed by planning margins should not be compared with capacity based on signalled headways – the utilisation would be misleadingly low.
Whichever method is adopted, some uncertainty will remain as to how many trains can be operated until a draft timetable has been prepared. Even then, the reliability of the operation remains in doubt until experience of operation is obtained. Short of this, benchmarking of utilisation statistics against current operations will at least indicate whether a similar level of reliability can be obtained. Even comprehensive simulations are best regarded as producing statistics for comparison of scenarios rather than being truly predictive in absolute terms.

Route Utilisation Strategies carried out by the UK’s Strategic Rail Authority have adopted a “compression” approach based on planning headways to compare capacity and usage, and inform decisions on capacity allocation. This is effectively capacity rationing by decision, and reflects the real situation facing the UK in the short and medium term, in that the network cannot provide for reliable operation of all the trains that franchised operators wish to run. It is an irony that one of the reasons for growth leading to capacity rationing in a supposedly-market environment is regulation of fares such as was never applied to the nationalised industry.

Further scope is seen for use of simulation to assess robustness of timetables, but will largely depend upon availability of simulation systems that are accessible to the staff of individual operators and can produce useful output with realistic timescales.

References