Determining Rail Network Accessibility

By Ben Waterson [(a)], Simon Box [(a)] and John Armstrong [(a,b)]

(a) Transportation Research Group, University of Southampton, (b) Arup

5th IMA Mathematics in Transport Conference, London, 12-14th April 2010

Abstract

The usual representation of optimal path finding problems within transport networks is focused on well established algorithms for identifying the optimal path (or set of paths) between two specific network nodes. When the required solution is the identification of the optimal route between every possible pair of nodes in the network however, these algorithms are inefficient.

The Floyd-Warshall algorithm provides an efficient way to compare all possible paths through each pair of nodes more efficiently, requiring only N^3 comparisons for a network of N nodes. To illustrate the potential of this approach to network analysis within transport research, this paper considers the issue of determining accessibility between railway stations (on the route between Weymouth and London Waterloo) served by a mixture of high-speed and stopping services.

A rail network is physically defined by the locations of tracks, but travel times are also dependent on whether stations are visited by high-speed services as well as stopping services. A single rail route therefore has to be represented not as a (topologically) straight line, but as a more traditional graph with high connectivity between nodes. Reformulating this into a matrix-based definition allows the Floyd-Warshall algorithm to efficiently determine the optimal routing (and hence travel times) between each pair of stations and therefore overall levels of accessibility to be determined.

1. Introduction

Continued investment in UK rail infrastructure (e.g. the West Coast Main Line project (SRA 2003) and High Speed Rail 1 and 2 (Butcher 2009, DfT 2009)) and timetable developments have enabled headline UK train speeds to increase to 200kmh^{-1} and European train speeds to reach 320kmh^{-1} (DfT 2009). This has contributed to a significant rise in rail patronage levels through both increased passenger numbers and increased distances being travelled. Considering travel on the rail network in more detail however, presents a slightly less homogeneous situation than the headline figures would suggest, with trains at the highest travel speeds only tending to serve a small subset of the total number of train stations, with travellers from the remainder relying on slower ‘stopping’ services to act as feeder trains into the high-speed subnet.

To truly understand typical travel times within a rail network therefore, it is necessary to understand not only the need to change trains when there is no direct service between stations, but also to cope with potential differing levels of service along the same sections of track. This paper therefore attempts to derive a methodology for determining rail network ‘accessibility’, based on averaging the minimum theoretical travel times from each station to all other stations in the rail network. By applying this methodology to the route between Weymouth and London Waterloo stations (Figure 1), the impact of high-speed and stopping services is clearly demonstrated.

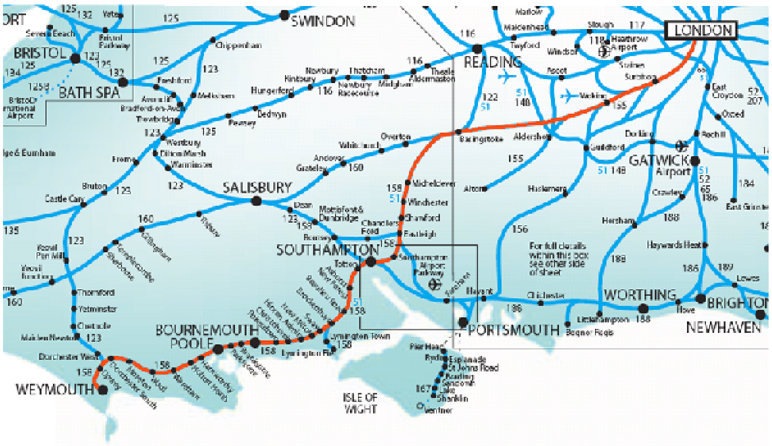

FIGURE 1. Weymouth to London Waterloo rail route (after Network Rail 2009)

2. Timetable Data

To ensure realistic distributions of both travel times and patterns of stations served by different stopping and high-speed train services, the base data for this paper is taken from the (December 2007) weekday timetabled passenger trains which stop at more than one of the fifty-four stations on the 235km 'South West Main Line' between Weymouth and London Waterloo. The base data is produced by identifying the minimum travel time for any train stopping at each pair of stations. This produces an asymmetrical travel time matrix as some pairs of stations are visited by trains travelling in one direction only (e.g. there exists a direct train from West Byfleet to Southampton Central, but not a direct train from Southampton Central to West Byfleet), and buffer times built into the timetables to account for anticipated delays on approach to busy stations (e.g. London Waterloo) lead to slightly higher travel times towards these stations than away from them.

It should be noted here that this data represents the minimum scheduled travel time by any single train between each pair of stations on the South West Main Line route. It does not therefore represent frequency of service (some pairs of stations may only be served by one train per day whereas other pairs often have multiple trains per hour) or efficiency in timetabling to ensure that passengers can make connections between trains with the minimum of delay. There may also be situations where train services exist between two stations on the South West Main Line route, but that the service does not solely use the direct route. The most noticeable effect of this is on Queenstown Road station which has direct connections to stations between Clapham Junction and Woking via other routes only. This results in minimum travel times to and from Queenstown Road being significantly longer than would otherwise be expected.

3. Graph Theory

The simplest method of determining accessibility within the rail network is to consider only the physical connectivity. Consider therefore the (topologically) linear graph (G_1) representing the railway line and set of stations (N) through Southampton (Figure 2), with the number of stations (nodes) $|N| = 6$ and allowing bi-directional travel along each edge. While being a connected graph is a prerequisite to functioning as a transport network, it is immediately clear that levels of connectivity are at the theoretical minimum, with node connectivity $K(G_1)$ = edge connectivity $K'(G_1) = 1$ and the maximum degree $\Delta(G_1) = 2$.

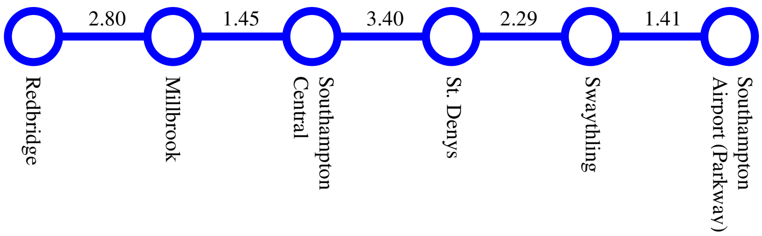

FIGURE 2. Distances (km) between Southampton stations

Considering only stopping services (i.e. those which do not pass through a station without allowing passengers to board/alight there) and therefore identifying the minimum travel time between adjacent (connected) pairs of stations provides the travel time matrix **T** given below, with $t_{ik} = t_{ij} + t_{jk}$ when station j is located between stations i and k in Figure 2.

$$\mathbf{T} = \begin{bmatrix} & 3 & 5 & 9 & 12 & 14 \\ 3 & & 2 & 6 & 9 & 11 \\ 5 & 2 & & 4 & 7 & 9 \\ 10 & 7 & 5 & & 3 & 5 \\ 13 & 10 & 8 & 3 & & 2 \\ 15 & 12 & 10 & 5 & 3 & \end{bmatrix} \begin{matrix} \text{Redbrdge} \\ \text{Millbrook} \\ \text{S.Central} \\ \text{St.Denys} \\ \text{Swaythg} \\ \text{S.Airport} \end{matrix}$$

Transforming **T** into estimates of accessibility for each station (**A**) can be done through averaging the reciprocals of travel times (3.1). The use of reciprocals allows for two issues to be addressed, firstly the need to allow for infinite travel times (in subsequent analysis where pairs of (non-adjacent) stations which are not served by a direct train service will be considered) and secondly that the benefit of a unit change in travel time perceptually tends to zero as the base travel time tends to infinity. Perhaps a more useful measure of accessibility however is 'relative accessibility' (**A′**) defined by (3.2), which enables a fairer comparison between networks of different geographic sizes (where stations in (geographically) larger networks will tend to have lower accessibility simply due to the larger travel distances involved).

$$a_i = \frac{\sum_{j \neq i} t_{ij}^{-1}}{N-1} \tag{3.1}$$

$$a_i' = \frac{\sum_{j \neq i} \left(\left(t_{ij} - c_i \right)^{-1} \right)}{\max_k \left(\sum_{j \neq k} \left(\left(t_{kj} - c_k \right)^{-1} \right) \right)} \tag{3.2}$$

It is worth noting however that defining accessibility (as here) based on travel times from a station may give a different picture to a definition based on travel times to the same station. For stations where there is a high level of demand compared to the number of available platforms (especially London Waterloo in this example), additional buffer times may be built into the timetable to protect against possible delays, but such times would not be added to services leaving the station. With the possible exception of this small group of very busy / bottleneck stations however, the differences in accessibility between journeys from and journeys to the station will likely be negligible and therefore journeys from the station will be used for the accessibility calculations in this paper.

Using (3.2) to calculate relative accessibility for the six Southampton stations (Figure 2) gives relative accessibility values of $\mathbf{A'}$ = (0.66, 1, 1, 0.81, 0.94, 0.79), suggesting as would be expected that the internal node stations have higher accessibility than end node stations. The lower relative accessibility for St. Denys station (0.81) is related to the higher distances from St. Denys to its adjacent stations compared to the other stations, which causes slightly higher travel times from St. Denys to Swaythling and Southampton Central and thus a lower relative accessibility. While realistic within the small number of stations being considered at this stage of the analysis however, this effect is comparatively less significant when applying the same methodology to the fifty-four stations on the full Weymouth to London Waterloo route.

4. Theoretical Frameworks

While the simplistic approach above of basing travel times on the cumulative travel times between adjacent pairs of stations clearly does not represent the true mix of train speeds and diversity of train services, it does illustrate that the 'between all station pairs' travel time matrix is the fundamental basis for the calculations. To derive a more realistic travel time matrix however, it is necessary to consider (firstly) the issue of high-speed services. These services can be represented within the graph as edges running parallel to the topologically linear representation used in Figure 2 above, with Figure 3 representing the typical situation on the subset of stations between Southampton and Winchester inclusive. It can be seen that while stopping services visit all stations on the route, high-speed services achieve lower travel times, but only visit a subset of stations (Southampton Central, Southampton Airport and Winchester in this example).

While this formulation allows for both high-speed and stopping train services however, it does not truly represent the situation faced by a traveller, because it does not differentiate between the time taken for a traveller to change trains at a station and the time taken for a traveller to simply enter the station and leave on the same train (a minimal time which is already accounted for in the timetable data and hence in the travel time estimations). Because only typical travel times are required for estimating accessibility, rather than specific timetabled differences in arrival and departure times, this changing trains delay can be represented as a penalty for each node (station) visited on the path of the form $\mathbf{C}$, where $c_i = fn$(size and complexity of station i) as larger multi-platform stations would typically require larger interchange times.

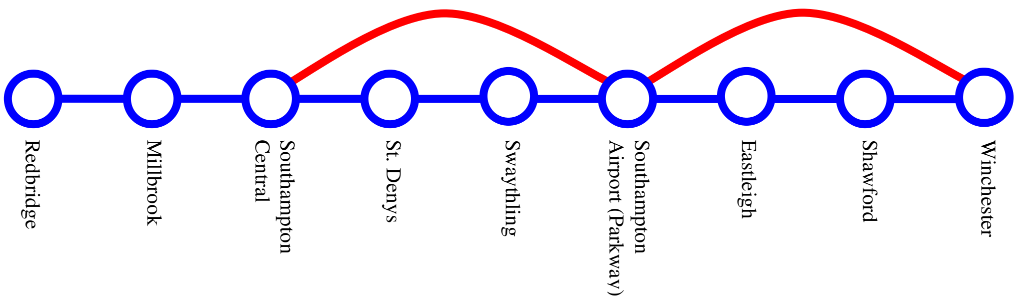

FIGURE 3. High-speed and stopping services between Southampton and Winchester

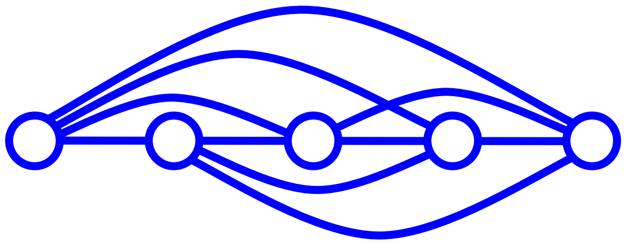

FIGURE 4. Fully connected graph of journey interchange points

Incorporating these penalties into the travel times by applying them to every edge leaving a node ($t'_{ij} = t_{ij} + c_i$) enables accessibility to still be calculated from a travel times matrix using (3.2), but this will lead to an additional penalty of c_i for all journeys starting at station i. While this may be true for a traveller in reality as few journeys truly start at the point of getting on the first train, the appropriate interchange cost (c_i) for the station of origin needs to be subtracted from the calculated overall journey travel times to ensure a fair accessibility comparison.

To ensure that the interchange penalties (**C**) are only applied to travellers who change trains at a station the network graph must be redrawn so that each edge represents a single train between interchange points (or journey origin / destination stations), regardless of any stations that it passes through en-route (whether it stops at them or not). For example see Figure 4, where the top edge represents a journey on one train between the end node stations, either a direct high-speed service or a stopping service if no high-speed service exists. Travellers using this edge of the graph would therefore not incur interchange penalties at the intermediate stations even if the train did physically stop there.

It should be noted here that for a single rail line with stopping services visiting all stations, this formulation will lead to a fully connected graph ($\forall i,j\ \exists\ t_{ij}$), but in general it may be necessary to add additional edges (with $t_{ij} = \infty$) for pairs of stations not served directly by any single train to achieve this state. The full formulation of the travel time matrix **T′** representing the interchange time and minimum travel time for all trains serving each pair of stations is therefore given by (4.1) with the relative accessibility calculated using (3.2) with values from **T′** substituted for those from **T**.

$$\mathbf{T'} : t'_{ij} = \begin{cases} 0 & \text{if } i = j \\ t_{ij} + c_i & \text{if } i \neq j \text{ and a train service from } i \text{ to } j \text{ exists} \\ \infty & \text{if } i \neq j \text{ and no train service from } i \text{ to } j \text{ exists} \end{cases} \quad (4.1)$$

4.1. Shortest Path Approach

Basing relative accessibility calculations on **T′** however implicitly assumes that travellers use only direct trains to travel between their origin and destination station, an invalid assumption as (a) journeys are possible in reality between pairs of stations not both served by a direct train and (b) even if a direct train between a pair of stations exists its travel time may not be optimal if it is a stopping service. This assumption is unnecessary however, as the change from a topologically linear to a fully connected graph structure now means that travellers effectively face a choice of paths through the graph, between which it is necessary to identify the minimum time route (usually referred to as the ‘shortest path’). Indeed the mixture of high-speed and stopping services implies that even an assumption of monotonic direction of travel cannot be assumed as it maybe advantageous between some station pairs to move counter to the prevailing (geographic) direction of travel to utilise stations served by a high-speed service (see Figure 5 - where [values] represent theoretical t'_{ij}).

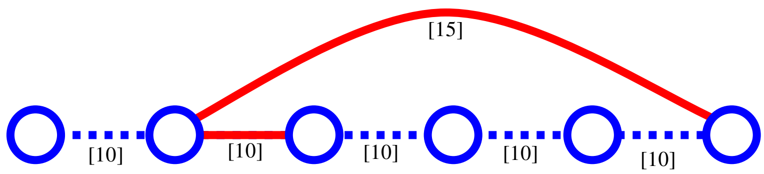

FIGURE 5. Reverse travel in optimal route

Solutions to the problem of finding the shortest path between a single pair of nodes are well established, with the commonest being Dijkstra's and the A* algorithms. While these algorithms could be applied in this situation to generate **T″** (a matrix of minimum travel times by any combination of train services), this would be computationally inefficient as the requirement here is to find the shortest path between all pairs of nodes. Taking Dijkstra's algorithm as an example, the worst-case performance for finding a single shortest-path is usually given as $O(|E| + |N| \ln|N|)$ where $|E|$ is the number of edges and $|N|$ the number of nodes (stations) in the graph (Cormen et al. 1991). In a fully connected network where $|E| = |N|^2$ this becomes $O(|N|(\ln|N|+|N|))$ and therefore applying the algorithm independently to $|N|(|N|-1)$ possible journeys gives a worst case in excess of $|N|^4$. While this worst-case could easily be improved on in practice (e.g. by allowing the algorithm to traverse all nodes within the graph from each starting node rather than terminating when a specific end node is reached, or reusing information about known shortest paths stored from earlier computations) this is still inefficient compared to algorithms designed specifically for solving all-pairs shortest path problems directly.

4.2. The Floyd-Warshall Algorithm

The widely established Floyd-Warshall algorithm (see Cormen et al. 1991 for example) is specifically designed to identify the shortest paths between all pairs of nodes in a fully connected graph with no negative cycles, with $O(|N|^3)$. This is precisely the situation with **T′** where all pairs of nodes are connected by a single edge (albeit potentially of infinite travel time) and all individual edge travel times must be positive (from (4.1) $\forall$ i,j $t_{ij} > 0$ if $i \neq j$ and $t_{ij} = 0$ if $i = j$). Rather than considering paths radiating from each individual node in turn, the Floyd-Warshall algorithm considers first all paths of length 1 edge (hence the need for a fully connected graph), then all paths of at most 2 edges etc. until all paths up to at most $|N|$-1 edges have been considered (the algorithm can terminate at this point as the absence of negative cycles means that paths containing more than $|N|$-1 edges cannot produce lower total travel times).

Drawing parallels to the situation faced by a traveller considering the quickest way to make a journey, the traveller would be first considering the time taken to use a single train from origin to destination, then consider all possible journeys including one change of train (at any other station in the network other than the start and end stations) to identify if any are quicker, then journeys containing two changes of train etc.. Because the Floyd-Warshall algorithm considers all possible k-1 edge journeys before considering k edge journeys however this simply reduces to calculating the (known) minimum k-1 edge journey to the final interchange station plus the additional time taken to travel from the final interchange node to the end station, and then selecting the minimum.

Formally, the Floyd-Warshall algorithm defines $\mathbf{T^{(k)}}$ to represent the minimum travel times between pairs of stations using at most k edges (trains), using $\mathbf{T^{(1)}} = \mathbf{T'}$ and calculating $\mathbf{T^{(k)}}$ for $2 \leq k \leq |N|-1$ through the dynamic programming formulation given in (4.2).

$$t_{ij}^{(k)} = \underset{p \in N}{\text{minimum}}\left(t_{ip}^{(k-1)} + t_{pj}^{(1)}\right) \qquad (4.2)$$

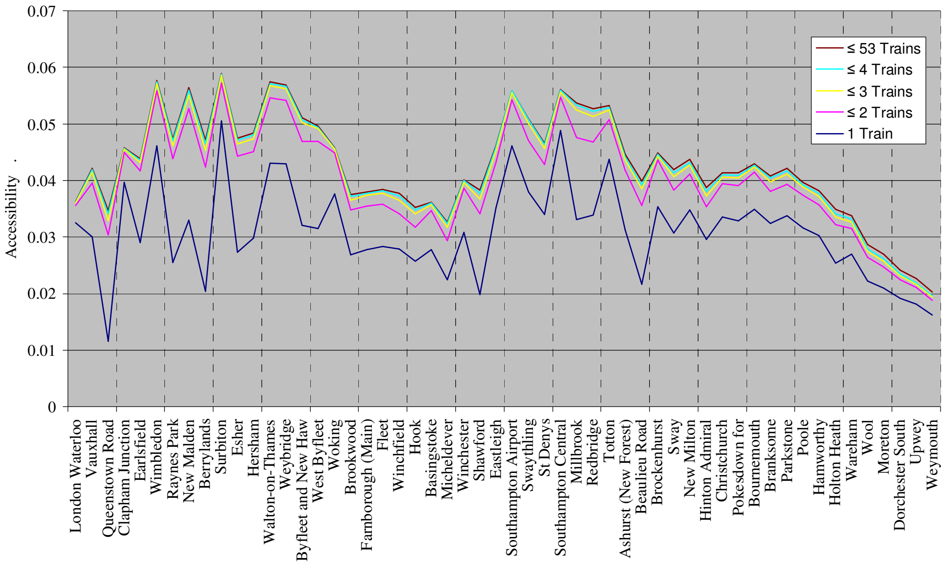

FIGURE 6. Station accessibility with **C** = 0

5. Results

Defining the basic (direct train) travel time matrix (**T′**) using (4.1) and applying the Floyd-Warshall algorithm in (4.2) therefore enables a calculation of the accessibility (**A**) and relative accessibility (**A′**) of the $|N| = 54$ stations on the South West Main Line between Weymouth and London Waterloo.

5.1. Zero Interchange Penalties

The first results to consider are those where $\forall$ i $c_i = 0$, i.e. where there is assumed to be no minimum interchange time at any station and therefore no penalty on a traveller for changing trains. Although unrealistic in practice this gives an upper bound on accessibility by representing the perfect timetabling (and station design) system. By considering different maximum values of k (the maximum number of different edges (trains) used to create the journey) these results (Figure 6) clearly show the impact of high-speed and stopping services. When only direct trains are considered (the lowest line in Figure 6) accessibility values (excluding Queenstown Road) fall in the range 0.016 (Weymouth) $\leq a_i \leq$ 0.051 (Surbiton), but as soon as a change of trains is allowed (k = 2) this range increases to $0.019 \leq a_i \leq 0.057$ as (a) travellers from stations only served by slower stopping trains can begin to take advantages of high-speed services for longer travel distances and (b) some journeys which were not possible by direct trains ($t^{(1)}_{ij} = \infty$) can now be completed.

Increasing the maximum number of trains to k = 3, means that almost all (longer distance) travel can now take advantage of faster trains (the range of accessibilities increasing slightly to $0.019 \leq a_i \leq 0.059$), with the benefit most noticeable at stations not served by high-speed services (e.g. Esher, Hersham, Walton-on-Thames, Weybridge, etc.) as k = 3 allows travel from these stations to other ‘non high-speed’ stations to utilise a high-speed service between two stopping services to achieve a lower travel time. Beyond k = 3 however improvements in accessibility values are small with there being little difference between k = 4 and k = 53 (the theoretical maximum for |N| = 54 stations), both giving ranges of accessibility of $0.020 \leq a_i \leq 0.059$.

Using the minimum possible travel times (i.e. k = 53) and plotting the relative accessibility (**A′**) rather than absolute accessibility (**A**) allows the overall situation to be examined. The general underlying pattern (Figure 7) follows the expected ‘dome’ shape with stations towards the middle of the route having higher accessibility than those close to the end nodes of Weymouth and London Waterloo, but two other effects are also noticeable.

The clearest deviation from the expected dome-shape pattern is the low relative accessibility of stations between Woking and Eastleigh. Being located towards the centre of the route these stations would be expected to have high relative accessibilities, but they actually all have values of $a'_i \leq 0.8$. To understand the reason for this it is necessary to consider the impact that using reciprocals has within the accessibility calculations. Reciprocals are used to recognise that the benefit of a unit change in travel time perceptually tends to zero as the base travel time tends to infinity, i.e. that differences between small travel times should have a greater impact than the same differences between large travel times. The stations between Woking and Eastleigh are located between the two dense clusters of stations around outer London and Southampton. This leads to them having greater travel times to adjacent stations than would likely have been the situation if the stations were more evenly spread. These comparatively high travel times do therefore represent a lower accessibility from these stations and this impact is therefore reflected in the accessibility calculations and results.

The second effect that can be identified in Figure 7 is the benefits to accessibility of high-speed services. Even allowing for travellers to change between stopping and high-speed trains, those stations served directly by high-speed services do have slightly higher relative accessibilities. This effect is most evident outside of London (for example the six stations marked with arrows in Figure 7), where stations are generally more widely spaced and hence the benefits of higher-speed travel are magnified.

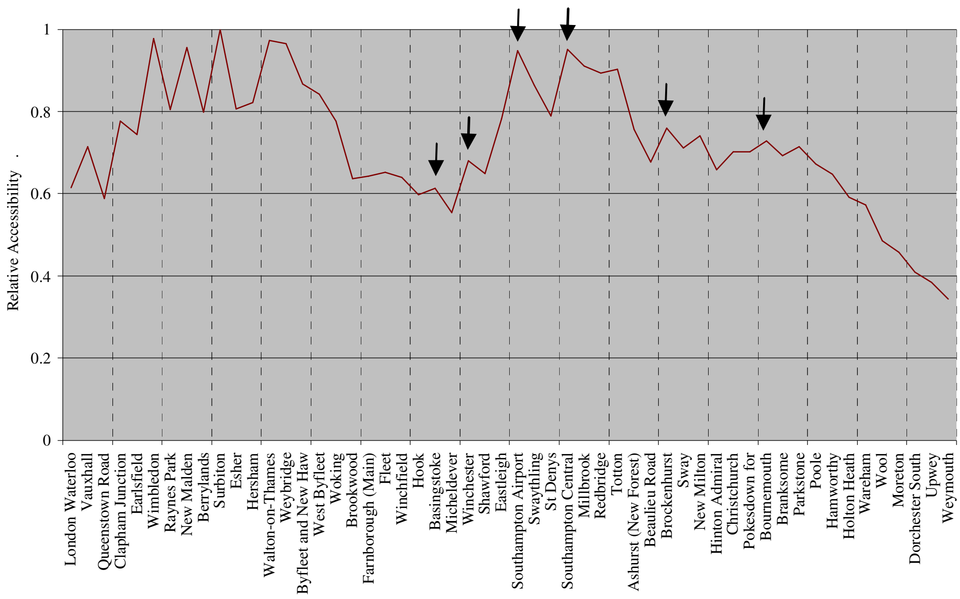

FIGURE 7. Station relative accessibility with **C** = 0

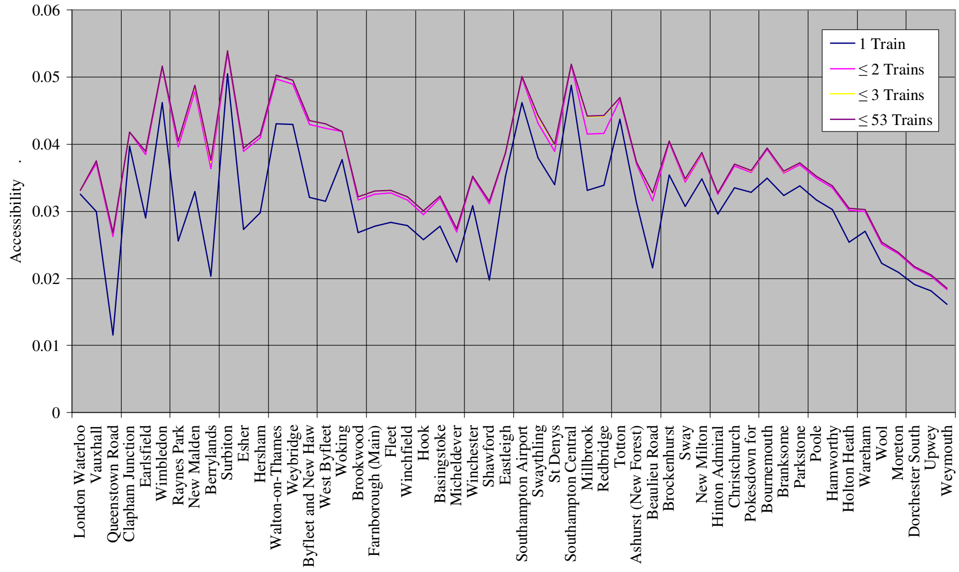

FIGURE 8. Station Accessibility with **C** = 5

5.1. Non-zero Interchange Penalties

While the inclusion of non-zero interchange penalties within the calculations does not affect the overall pattern of accessibility (Figure 8), it does have an impact on the choices of paths through the graph and the resulting overall accessibility values. The clearest consequence of this is that fewer individual trains are used to produce the optimal journeys, with the small reductions in travel times which were being achieved by adding (for example) a 4th or 5th separate train into a journey (Figure 6) not outweighing the interchange cost incurred by doing so. This result actually illustrates that the underlying timetable for the route is essentially consistent, with most optimal journeys consisting of at most three trains (e.g. a stopping service, high-speed service and stopping service) rather than it being possible to achieve significant travel time savings by making multiple changes en-route.

6. Conclusions

This paper has therefore revealed that the ideas of accessibility which are normally applied to comparatively high connectivity road systems can be equally applied to a linear rail route. By understanding that a mixture of high-speed and stopping services (and the penalties associated with travellers needing to change between services) can be represented using a more connected graph structure, the problem of finding the minimum travel time between any two pairs of stations has been shown to be equivalent to the traditional all-pairs shortest path problem for which efficient solution algorithms exist.

Applying one of these algorithms (Floyd-Warshall) to the transformation of passenger timetable data from the South West Main Line between London Waterloo and Weymouth has enabled the impacts of high-speed services, station density and interchange penalties to be combined. This has shown that stations in the middle of the route (between Woking and Eastleigh) have significantly lower accessibility than would have originally been anticipated. Identifying the relative accessibility of stations in this way provides a simple mechanism for rail operators to understand both the current situation and how planned timetable changes will have either a positive or negative impact on travelling from different stations.

To achieve maximum benefit this analysis needs to be extended beyond a topologically linear route to include multiple connected routes (Figure 1) and allow comparisons between peak, off-peak and weekend travel. Achieving this however is simply a matter of processing the data to create the direct trains travel time matrix ($\mathbf{T}'$), as although the methodology presented here was developed for a linear route and weekday services, no changes are necessary to relax these conditions and enable analysis of more complex situations.

REFERENCES

BUTCHER, L. 2009. Channel Tunnel Rail Link (HS1). *House of Commons Library, Standard Note SN/BT/267.*

CORMEN, T.H, LEISERSON, C.E. & RIVEST, R.L. 1991. Introduction to Algorithms, *MIT Press*.

DFT 2009. Britain's Transport Infrastructure: High Speed Two, *Department for Transport report.*

NETWORK RAIL 2009. National Rail Timetable Map 2009, *http://www.nationalrail.co.uk/system/galleries/download/print_maps/uk.pdf (accessed February 2010).*

SRA 2003. West Coast Main Line Strategy, *Strategic Rail Authority report (available February 2010 from http://www.dft.gov.uk/pgr/rail/pi/wcml/westcoastmainlinestrategy.pdf).*