Optimal timetabling with connections in underground railway systems

R. Minciardi[a], M. Paolucci[a], R. Pesenti[b]

[a]Department of Communications, Computer and System Sciences, University of Genova, Via Opera Pia 13, I-16145, Italy

[b]Department of Electrical, Electronic Engineering, and Computer Science, University of Trieste, Via Valerio 10, I-34127, Italy

Abstract

This paper investigates the problem of defining a timetable for the train runs in an underground railway network. The objective is to optimise the quality of the service offered to the customers. To this end a problem decomposition approach which favours the connections among the different lines is introduced.

1 Introduction

The problem of producing daily timetables (also called train run schedules) for underground railway systems, generally involving several lines, is dealt with in this paper. Such a problem may be formalised by taking into account different objectives and constraints. In particular, the possible objectives can be partitioned into two separate classes: a first class involving objectives relevant to the operational costs, and a second class of objectives relevant to the satisfaction of the system customers. On the other hand, the problem constraints take into account the operational conditions for the railway network under concern, as they can involve the size of the train fleet, the number of employees and their shifts, the power supply limits, and so on. Then, it is easy to recognise the definition of the train run schedule as a complex multiobjective problem.

This paper is focused on some specific aspects of the general scheduling problem, in particular, relevant to the quality of the service offered to the customers. A sensible way to evaluate such a quality can be simply to compute the average time spent by the customers into the system, that is, the average time spent by the passengers in the transportation system to complete their trips. Such a performance measure takes into account both the total travelling time and the total waiting time spent by the passengers. This fact is particularly important in multiple line railway systems, as the passengers could require to board more than a single run to reach their destinations, and consequently the total time they spend in the system can be heavily influenced by their waiting times at the various connections. Clearly, in order to compute the passenger average time in the system, the demand for service of the passengers, namely the passenger arrival rates organised according to their origins and destinations, should be known. The "transfer coordination" problem, namely the problem of defining a schedule trying to guarantee the coordination among runs on different lines, has been faced in the context of bus scheduling through two main approaches. The "timed transfer" approach prescribes that the runs should synchronise at specific connecting points, and it is applied to network with a simple structure such as a main/feeder lines system (e.g., see Salzborn [1]). Usually with such an approach several simplifying assumptions should be introduced, such as to impose constant headways for the lines, or to take into account only a few connecting points. On the other hand, the "transfer optimisation" approach is based on the optimisation of an objective function which penalises the lack of synchronisation among directly connected lines (e.g., see Bookbinder et al. [2]). It should be pointed out that, in the case of underground railway networks, which are usually high frequency transportation systems, the transfer coordination is worth when the involved headways are sufficiently long (see Knopper [3]), as it may happen during the off-peak periods. However, the results proposed in this paper could present a general interest as they can be easily generalised to other transportation systems, as long distance railways and buses.

The work that is presented in this paper follows the transfer optimisation approach, even if with a slightly different rationale. Clearly the formulation and the solution of the whole scheduling problem, which takes into account all the connections at the same time, is a very difficult task, since, generally, the performance of a schedule for a single line may be influenced by the schedules of other connected lines and vice versa. Then, a procedure based on a problem decomposition is proposed which first sorts the lines according to a given criterion, and iterates defining the schedule for a single line at a time by means of a scheduling policy, presented by Minciardi et al. [4], which dispatches the train runs for single line as a function of the passenger arrival rates at the stations. Hence, in order to try to force the single line policy to synchronise the runs of a not yet scheduled line with the runs on already scheduled ones, at each step of the above procedure the passenger arrival rates for the line under concern should be properly modified. In such a way, the proposed procedure tries to determine a compromise among the possible contrasting service requirements of the passengers of a line, as the departure times of the runs are fixed depending on the relative amount of the passengers bound for different destinations on the line, and, among them, on the passengers requiring connections with other already scheduled lines.

In the following section the model of the network is presented, then the multiple line scheduling procedure is introduced, and, finally, an example of its application is proposed.

2 The railway network model

The underground railway network considered in this paper is composed by several loop lines which are not physically connected, namely they neither share tracks, nor present crossings. However, the system connectivity is guaranteed by the possible synchronisation of train runs at connecting stations.

Let us consider, for example, a railway network in which U single lines are present. Each single loop line joins two terminal stations with a double track. A line u, for u=1,...,U, connects N_u platforms identified by the couple (k,u), for k=1,...,N_u-1. The stations of the network are sets of pair of platforms, and the connecting stations include a pair of platforms for each line they connect. For each line u, the platform with index 1 corresponds to the first (main) terminal station, I_u is the maximum number of runs operated in a day, and M_u is the size of the train fleet. It is assumed that the trains cannot overtake each other while running on the line. In addition, the model here considered is entirely deterministic. Assuming that the trains move between the platforms always at a nominal speed a-priori fixed, and that they stop at the platforms for predefined dwell times, the only decisional variables that should be taken into consideration are the train run departure times from the main terminal station of a line u, i.e., the variables $t_{(i,u)}$ for each run (i,u), for i=1,...,I_u. Then, the following inequalities should be satisfied

$$t_{(i+1,u)} \geq t_{(i,u)} + h \qquad i=1,...,I_u-1 \qquad (1)$$

$$t_{(i+M_u,u)} \geq t_{(i,u)} + H \qquad i=1,...,I_u-M_u \qquad (2)$$

In inequalities (1), h represents the minimum headway between two successive runs such that they can keep the nominal speed. Hence, inequalities (1) prevent the safety control system (SCS) (see Gray [5]) from slowing down the train speeds in order to keep them at a safe distance. Inequalities (2) impose that two runs, (i,u) and (i+M_u,u), operated by the same train, should respect the minimum operating time H for a round trip, including the reversing time. It should be noted that the model introduced so far may be generalised by allowing as decisional variables the departure times from each platforms. However, the worth of such a generalisation is questionable, as it provokes a increase of the computational burden without a significant improvement in the schedule performance (see Minciardi et al. [4]).

As the scheduling objective is the optimisation of the service quality, a set of functions should be introduced to model the passenger arrival process at the different platforms. Let $n_{(k,u),(r,v)}(t)$ the arrival rate at platform k of the line u of the passengers bound for platform r of the line v, at time t. Obviously, u and v may be different. The objective relevant to passenger comfort can be evaluated by the following expression (3) representing the average time spent by the passengers in the system (ATS), i.e., the amount of time that, on the average, a passenger spends from the instant he/she arrives at the platform to the instant he/she gets to the destination. The index c((i,u),(r,v)) in eqn (3) corresponds to the run which stops at the platform (r,v) having the "best" connection (possibly through other intermediate runs) with the run (i,u).

$$\text{ATS}= \tag{3}$$

$$\frac{1}{Q}\sum_{u=1}^{U}\sum_{v=1}^{U}\sum_{i=1}^{I_u}\sum_{k=1}^{N_u}\sum_{r=1}^{N_v}\int_{t_{(i-1,u)}+p_{(k,u)}}^{t_{(i,u)}+p_{(k,u)}} n_{(k,u),(r,v)}(t)(t_{(c((i,u),(r,v)),v)}+p_{(r,v)}-t)dt$$

Note that the run c((i,u),(r,v))) coincides with (i,u) if u is equal to v, and that a positive passenger arrival rate may be associated only to a subset of all the possible pair of origin and destination platforms. The quantity $p_{(r,v)}$ is the operation time, assumed constant, taken by a train from the terminal platform (1,v) to the departure from platform (r,v). Hence, $t_{(i,u)}+p_{(k,u)}$ is the departure time from platform k, which in the following may be also indicated by $\tau_{i,(k,u)}$, and $t_{(c((i,u),(r,v)),v)}+p_{(r,v)}-t$ is the time between the arrival at the platform k at time t of a passenger and the arrival of its final train to the destination platform r. Finally, $t_{(0,u)}$ is defined as the opening time of the line u, and Q is the total number of passenger served during the operating time interval, namely

$$Q=\sum_{u=1}^{U}\sum_{v=1}^{U}\sum_{i=1}^{I_u}\sum_{k=1}^{N_u}\sum_{r=1}^{N_v}\int_{t_{(i-1,u)}+p_{(k,u)}}^{t_{(i,u)}+p_{(k,u)}} n_{(k,u),(r,v)}(t)dt \tag{4}$$

The function c(.,.) introduces a difficulty in the minimisation of ATS, as it is generally impossible to a-priori (i.e., before the complete definition of a timetable) associate each run (i,u) with its correspondent run c((i,u),(r,v)). For this reason, the approach here proposed decomposes the problem in a set of single line problems in order to find a feasible solution, hopefully close to an optimal one.

3 The multiple line scheduling approach

The rationale on which is based the decomposition approach adopted for the multiple line scheduling is to sort the lines according to an a-priori fixed importance criterion, and to schedule one line at a time assuming that the schedules of lesser important lines (secondary lines) could be adapted to the schedules of the more important ones (primary lines). In particular, the secondary lines are scheduled taking into account the following hypotheses:

1. the passengers from already scheduled lines arrive at platforms according to the train timetables of the origin lines;
2. the passengers, who have to continue their travel on at least an already scheduled line, arrive at the platforms on a secondary line as the runs coordinated with the desired runs on the next scheduled line were available.

According to such hypotheses, which may be justified by the observation of the actual behaviour of the passengers of a transportation system, the passenger arrivals to a secondary line can be modified in order to force the single line policy to generate a schedule synchronised with already scheduled lines.

3.1 The structure of the procedure

The structure of the multiple line scheduling procedure is the following:

```
Multiple Line Scheduling Procedure
        define a complete order relation among the lines of the network
        while there are not scheduled lines
                begin
                select the first line not scheduled yet
                recompute the passenger arrival rate to the line platforms
                generate a schedule for the line
                end
```

There are two critical phases in the above procedure: the sorting phase, as different sorting criteria may lead to different schedules (the above procedure clearly penalises passengers starting their trips on secondary lines); the arrival rate recomputation phase, as the train synchronisation is due to the capability of imposing peaks of demand for service at specific time instants. On the other hand, let us assume for now to be able to efficiently solve the single line problem.

3.2 The passenger arrival rates recomputation

To better understand this phase, the following assumptions must be introduced.

Assumption 1. No passenger trips include two or more times the same line.

Assumption 2. Each pair of origin and destination platforms is join by a single a-priori fixed "preferential" path. This path is the only one used by passengers in their trips from that origin to that destination.

Before scheduling a secondary line, for each connecting platform C on such a line, the arrival rate of passengers coming from any other platform P belonging to a different line, and bound for any other platform D, and such that C is on a preferential path (P,D) starting from P and ending in D, is recomputed as follows:

- if the (P,D) includes only not yet scheduled lines, the arrival rate from P to D is used to compute an internal arrival rate to platform C of passengers from P and bound for D. Such a rate is obtained recursively by computing analogous arrival rates to possible intermediate connecting platforms. For any pair of successive connecting platforms, the internal arrival rate to the second platform is defined by means of a time shift applied to the arrival rate to the first platform. Such a shift is equal to a quantity, here called "expected nominal travelling time" (ENTT), which corresponds to the a-priori determined nominal time spent by a passenger travelling between the two platforms, taking into account both the nominal travelling and waiting times;
- if (P,D) includes already scheduled lines, define W the last connecting platform on the last, in the topological sense, scheduled line g along (P,D). Then, it reasonable to assume that the passengers from P want to arrive at W just in

time to board the runs already scheduled on g, in order to reduce their waiting at W. Hence, defining $\tau_{i,W}$ as the departure times of the runs i from W, the passengers should arrive at W at the time instants $\tau_{i,W}-\varepsilon$, where ε, call it "earliness factor", is an opportune time interval a-priori fixed. For each pair of successive connecting platforms, (A,B), which precede W along (P,D), the time instants of the desired arrivals at the first of the connecting platforms, A, are backward recursively defined, starting from time instants $\tau_{i,W}-\varepsilon$, according to the following rules:

- if the line used to move from A to B has not been scheduled yet, the desired arrival times at A are obtained shifting back the correspondent ones at B of the ENTT between A and B;
- otherwise, the desired arrival times at A are obtained from the correspondent ones at B, determining, for each of these last ones, the departure time from A of the last scheduled run allowing the passengers to reach B in due time, and subtracting from it the earliness factor.

The above recursion ends at the departing platform P. If C precedes W on (P,D), then define $b_{i,(P,D)}$ the desired arrival times at C, computed by the previous backward recursive procedure, corresponding to the instants $\tau_{i,W}-\varepsilon$. Then, the arrival rate at P bound for D is redefined as a sequence of impulses which occur at the desired arrival instants at P computed as above. Each impulse represents the group of passengers who originally have entered the line from the outside bounded for D between the time of occurrence of the impulse itself and its predecessor. Finally, the new rate redefined in this way is used to recompute the correspondent rate of passengers from P to D at platform C. This last rate is obtained by a forward recursive procedure for each pair of successive connecting platforms, (A,B), preceding C along (P,D). If the line between A and B has not been scheduled yet, the rate in B is computed by shifting forward of the ENTT the impulsive rate in A. If the line between A and B has been scheduled, the rate in B is defined by the arrival times of the runs from A that serve the impulsive arrival rate of passengers coming from P. This recursive phase stop in C. Then, define $f_{i,(P,D)}$ the time instants at which the arrival impulses occur at the platform C, and, if the corresponding $b_{i,(P,D)}$ have been not previously defined, let $b_{i,(P,D)}$ be equal to $f_{i,(P,D)}$.

Some further notation should be introduced before more formally presenting the passenger arrival rate recomputation procedure. Consider a pair (A,B) of successive connecting platforms. Let F(t,A,B) be the ENTT between A and B, if the connecting line is not scheduled, otherwise be equal to the time required to arrive at B from A by the passengers who get the first train departing from A after or at instant t. Let G(t,A,B) be equal to F(t,A,B), if the line connecting A and B is not scheduled, otherwise be equal to the time required to arrive at B from A by the passengers who get the last run from A which make they able to arrive at B before or at instant t. Let r(j) a function that returns the j-th connecting platform on (P,D). If there are n of such connecting platforms between P and D, then r(1) = P, and r(n) = D.

```
passenger_arrival_rates_recomputation (line g, connecting platform C)
begin
for all preferential path (P,D) including C
   begin
   n is the number of connecting platforms in (P,D);
   $k=r^{-1}(C)$; $m=r^{-1}(W)$
   if there are not scheduled lines in (P,D)
        for j = 2 to k
                 $n_{r(j),D}(t) = n_{r(j-1),D}(t - F(t,r(j),r(j-1)))$;
   else
        begin
        define t(i,j) the desired arrival time at r(j) to be in time for run i at W;
        for all i runs leaving from W
                begin
                $t(i,m) = \tau_{i,W} - \varepsilon$
                for j=m-1 down to 1
                      $t(i,j) = t(i,j+1) - G(t(i, j+1), r(j), r(j+1))) - \varepsilon$

                      $n_{P,D}(t) = \sum_i \delta(t - t(i,1)) \int_{t(i-1,1)}^{t(i,1)} n_{P,D}(t)dt$

                for j = 2 to k
                      $n_{r(j),D}(t) = n_{r(j-1),D}(t - F(t,r(j),r(j-1)))$;
                end
        end
   end
```

Note that the arrival rates of passengers travelling only on single lines are never modified by the above procedure. Furthermore, it is obvious that the results of the procedure depend on the value fixed for the earliness factor, and hence it may be sensible to perform a tuning phase aiming at determining the optimal value for ε. Finally, note that for each preferential path (P,D) including C a set of time intervals $[f_{i,(P,D)}, b_{i,(P,D)}]$, here also called "service windows", is defined. In case that the ENTTs on the not scheduled lines are not exceeded, the departure of runs at any instant in such intervals allow the passengers arrived at C at instant $f_{i,(P,D)}$ to get the subsequent connection on (P,D) departing at time $t(i,k+1)+ \varepsilon$. For such a reason, and for the way a single line is scheduled on the basis of the passenger arrival rate, as it will be shown in section 3.3, the arrival rate of passengers from P at C could be furtherly redefined, if convenient, by shifting the time occurrence of the arrival impulse from $f_{i,(P,D)}$ to any instant in the interval $[f_{i,(P,D)}, b_{i,(P,D)}]$. In particular, a possible choice is to initially determine, for each connection C, all the intersections among the service windows associated with all the possible preferential paths including C. Such intersections define the time intervals in which the departure of a run satisfies the connection requirements of passengers following different preferential paths. Then, it may be convenient to assume that the passengers who can be served by the same run in one particular intersection of service windows, arrive contemporary at a time instant in such an intersection. In general, a service window $[f_{i,(P,D)},b_{i,(P,D)}]$ can include more disjoint intersection intervals, and then it should be decided to which of such intervals the passengers arriving in the service window from P to D are assigned. If those passengers are assigned to the first, from a temporal point of view, intersection interval, then they are more likely, even in the relaxed hypotheses for the not already scheduled lines, to be able to board the desired connecting runs. On the other hand, if the last intersection interval is chosen, in the same relaxed hypotheses the passengers are more likely to board the desired connecting runs even in presence of travelling time longer than the ENTT to reach the platform C.

3.3 The single line scheduling

The basic result is relevant to single line scheduling problem, given the passenger arrival rate, can be found in Newell [6]. Unfortunately, such a result cannot be used in a single line scheduling subroutine for the procedure in subsection 3.2, as it cannot be applied in presence of impulsive arrival rates. For this reason, Minciardi et al. in [4] proposed new single line scheduling policies capable of dealing with such a kind of rates. In particular, the run departure times from the main terminal platform of a line u are determined by the instants $t_{(i,u)}$ satisfying equations

$$\int_{t_{(i,u)}}^{t_{(i+1,u)}} n(t)dt = \frac{2K}{h(t_{(i+1,u)} - t_{(i,u)})} \qquad i=0,\ldots, I_u\text{-}1 \tag{4}$$

or

$$h\int_{t_{(i,u)}}^{t_{(i+1,u)}} n(t)(t_{(i+1,u)} - t)dt = K \qquad i=0,\ldots, I_u\text{-}1 \tag{5}$$

where $n(t) = \sum_{k=1}^{N_u-1} \sum_{r=k+1}^{N_u} n_{(k,u),(r,u)}(t - p(k,u))$ is the sum of the arrival rates at the platforms of u, actualised to the main terminal platform. Clearly, for a line in a network, the contributions from both the outside and the connected lines are included in the arrival rates of the connecting platforms. The constant K is equal to the fixed cost for one run, and h to the passenger waiting cost for unit of time.

In policies (4) and (5) the dispatching of runs does not depend on the number of waiting passengers but on their cumulative waiting time: even a small number of passengers are served by these policies if they accept to wait a sufficient amount of time. Such policies, developed for impulsive n(t), provide the same results of Newell [6] for smooth arrival rates. However, (4) and (5), in some cases, can lead to schedules which make the passengers wait more than strictly necessary. A further policy, called NSP (Non-Smooth arrivals Policy), proposed by Minciardi et al. [4] overcomes the drawback of (4) and (5). The NSP is based on the observation that, in case of a rapidly varying n(t), the optimal schedule to make the runs serve the passengers in the decreasing phase of the n(t), in order to make most of the passengers arriving in that interval able to board the trains. Then, NSP tries to fix run departure instants, within a considered time interval, in correspondence with the n(t) decreasing phase. Policy NSP is composed by the three following steps:

1. given t_i, fix the instant, $t_{i+1,max}$, within which run i+1 should depart, e.g., taking the minimum between the values of t_{i+1} provided by (4) and (5);
2. compute the mean passenger arrival time, τ, and its standard deviation, σ, in the interval $[t_i,t_{i+1,max}]$;
3. select $t_{i+1}=\min[t_{i+1,max}, \tau+\alpha\sigma]$ con $\alpha\geq 0$;

Note that, if $\alpha = \sqrt{3}$, for constant n(t) policy NSP gives the same results of policy (4) or (5). In addition, when, in the considered time interval, the n(t) function is unimodal and symmetric, τ coincides with the local maximum of n(t) and then $\tau+\alpha\sigma$ is located in the decreasing phase of n(t), so that a percentage of more than $1-1/\alpha^2$ of the passengers arrived in such an interval is able to board the train.

4 An example

Consider now an elementary example of passenger arrival rate recomputation in a simple system including three lines as depicted in figure 1.

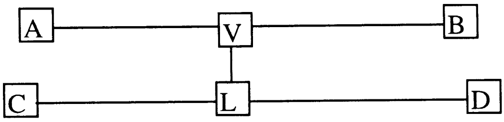

Figure 1: A simple three line system

Assume that the lines which connect A with B, and C with D have already been scheduled, and that only passengers who use the line connecting V with L are the ones moving between the other two lines. The problem is to schedule the line connecting V with L. These last stations are both the terminals of the line and the connections for the two other lines. Let L be the main terminal station. Assume that the headway between the runs on the scheduled lines is equal to one hour, and that each run from A to B stops at V at 28 minutes of an hour, and from B to A at 25 minutes of an hour. Each run from C to D stops at L at 12 of an hour and from D to C at 58 of an hour. Finally, the a-priori ENTT from V to L (and vice versa) is equal to 2 hours and 9 minutes. Due to the periodicy of the schedules already defined, from now on only the minutes will be reported when this is not cause of confusion, and the index denoting the runs will be omitted.

When scheduling the line from L to V and vice versa, it is convenient to express all the time instants referring to the main terminal L; then, for example, the arrival times at V of runs from A to B is 19, and from B to A is 16. By means of the passenger arrival rate recomputation procedure, the following set of service windows associated with the connection L can be determined:

Transactions on the Built Environment vol 18, © 1996 WIT Press, www.witpress.com, ISSN 1743-3509

$[f_{A,C},b_{A,C}]=[19,40]$ $[f_{A,D},b_{A,D}]=[19,54]$ $[f_{C,B},b_{C,B}]=[12,1h10]$ $[f_{C,A},b_{C,A}]=[12,1h7]$
$[f_{B,C},b_{B,C}]=[16,40]$ $[f_{B,D},b_{B,D}]=[16,54]$ $[f_{D,B},b_{D,B}]=[58,1h10]$ $[f_{D,A},b_{D,A}]=[58,1h7]$
As an example, the first service window indicates that a run should depart from L not before 19 to board at V the passengers bound for C just arriving from A at 28, and not after 40 to allow the above passengers to board at V not after 49 and then to get the connecting run from D to C which stops at L at 58. Figure 2 is a Gantt chart for a period of a hour representing the above service windows.

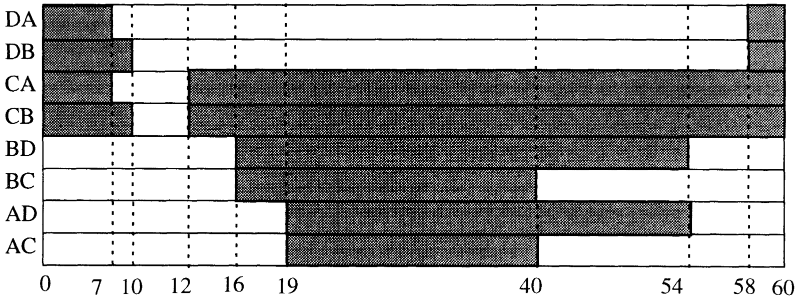

Figure 2: The chart of the service windows at L.

In figure 2 no common intersection exists among all the service windows. However, all the passengers requiring connections, except the ones travelling from D, can be served by a run departing from L at any instant between 19 and 40. A possible solution able to serve all the passengers can be found by modifying the schedule for the line between C and D (forcing the satisfaction of conditions in Salzborn [1]), in particular, anticipating to 40 instead of 58 the arrival of a run at L from D, and then imposing a dwell time of 18 minutes. An alternative, which also reduces the global waiting time, is that of delaying the departure from V to A of 12 minutes and from V to B of 9 minutes, since in such a way $b_{C,A}=b_{C,B}=19$ and $b_{D,A}=b_{D,B}=1:19$, and then fixing at 19 of an hour the run departures from L.

References

1. Salzborn, F.J.M. Scheduling bus systems with interchange, *Transp. Science*, 1980, **14**, 211-231.
2. Bookbinder, J.H., Desilets, A. Transfer Optimization in a Transit Network, *Transp. Science*, 1992, **26**, 107-118.
3. Knopper, P. Muller, T. Optimized Transfer Opportunities in Public Transport, *Transp. Science*, 1995, **29**, 101-105.
4. Minciardi, R. Paolucci, M. Pesenti, R. Generating optimal schedules for an underground railway line, pp. 4082 to 4085, *Proc. of 34th CDC*, New Orleans (LA), 1995.
5. Gray, B. H. *Urban Public Transportation Glossary*, Transportation Research Board, National Research Council, Washington, D.C., 1989.
6. Newell, G.F. Dispatching Policies for a Transportation Route, *Transp. Science*, 1971, **5**, 91-105.