STOCHASTIC APPROXIMATION TO THE EFFECTS OF HEADWAYS ON KNOCK-ON DELAYS OF TRAINS

MALACHY CAREY and ANDRZEJ KWIECINSKI
Faculty of Business and Management, University of Ulster, Northern Ireland, BT37 0QB

(Received 30 November 1992; in revised form 28 July 1993)

Abstract - In train planning and timetabling, the trip time on each link is assumed to depend on the type of train and characteristics of the link, but is usually treated as independent of the time interval (headway) between trains. However, in practice trains are subject to delays from a variety of causes, and since normally they are not allowed to pass each other on a link, any delay to one train may cause "knock-on" delays to following trains. This is especially true of the high density double and multiple track railways in Britain and Europe. The shorter the scheduled headway between trains, the greater is the expected knock-on delay and hence the greater the expected trip times of following trains. We develop simple stochastic approximations to these knock-on train delays. To test and calibrate the approximations, we conduct detailed stochastic simulation of the interaction between trains as they traverse sections of the link. The approximate relationships that we derive between scheduled headways and knock-on delays can be used, for example, (a) to provide correction factors for other stochastic or deterministic models of train planning, dispatching, or control; (b) to adjust train timetables, which are currently produced without explicitly considering the expected knock-on effects, and (c) to make it feasible to conduct larger scale simulations of train networks, by reducing or removing the need to simulate behaviour within each link.

1. INTRODUCTION

There is a substantial literature on modelling the distribution of trip times (delays) on rail lines. This literature is concerned with single track, or partially double track, lines (see surveys in Assad, 1980a, 1980b; Crainic and Queirin, 1987; and also Petersen and Taylor, 1982, 1987; Petersen, 1984; Chen and Harker, 1990). The delays occur since train meet-passing or overtaking cannot take place on the single track segments. For two trains to pass or overtake one train has to wait temporarily in sidings, which are located at intervals along the single track.

However, the types of train delay with which we are concerned in this paper are different from the above. The single track or partially double track lines with which the above authors are concerned are typical in North America and in some other countries. However, such lines are the exception in Britain and Europe. There, typically (a) there are two or more lines (usually four) between each pair of train stations, (b) each line is dedicated to trains in one direction (e.g., two lines up and two down), (c) train meet-passing problems do not arise as trains in opposite directions use different lines, and (d) overtaking takes place only at station stops (e.g., see Carey and Lockwood, 1992; Carey, 1993). We here consider such a line with trains travelling in one direction. The scheduled speed can be different for each train and can vary along the line. Unscheduled delays occur due to random variation in train speeds, which in turn can delay following trains. Delays to following trains also depend on the train control and signalling system, the required minimum headway between trains, and speed restrictions on following trains. We use the term knock-on delay to refer to that part of a train's delay which is caused by other trains in front of it.

In train planning it would be very useful to have even approximate predictions of how the distributions of trip times of trains would be affected by increasing or decreasing train headways and hence knock-on delays. It would be especially useful to have prior

1At present at the Mathematical Institute, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland.
estimates of parameters such as the mean, 90th and 95th percentile, etc., of the resulting link trip time distributions. These parameters are important as performance indicators in train planning: train planners usually wish to, or are required to, set link trip times which will be sufficient to ensure punctuality on 90% or 95%, etc., of trips. In Britain and some other countries, rail operators are required to collect and publish such performance statistics for each line. However, due to difficulties in predicting these knock-on delays, train planners commonly ignore them, or greatly overestimate or underestimate them at the planning stage.

Exact analytic expressions for these parameters are difficult or impossible to obtain, as they depend on train interaction and signalling systems over the many segments within each link (see below). An alternative approach is to perform detailed simulations of train running. However, such large scale simulations are too costly or time consuming to undertake for each proposed adjustment of the train plan. Such simulations would require detailed data on the characteristics of each segment of the line (including the signalling system, trip time distributions for each type of train, etc.), and would require setting up an explicit simulation model of these. Also, to obtain results that are consistent in repeated simulations, the size of the (pseudo-random) sample for each train in the simulation should grow roughly in proportion to the number of signals, crossings, etc., on the line. Further, a single line or link is usually only a small component in a larger rail network. If every such line required a detailed simulation, then a simulation of the whole network may become almost intractable. We are therefore concerned here to find simpler approximate methods of estimating parameters of link trip time distributions, including knock-on delays. This would reduce the need for detailed simulation of the activity along each line.

In Section 2 we set out approximation equations which are to be tested in the rest of the paper. These relate the mean and xth percentiles of train delays to the headways between trains. We can assume that trip time distributions in the absence of knock-on effect (e.g., when headways are much larger than required) are available or can be estimated. In Section 3 we state algebraically the “2-aspect signalling” and “3-aspect signalling” systems which control train behaviour on a link, and relate these to the approximation equations of Section 2. In Section 4 we report on a set of simulation experiments that we carried out to test the accuracy of the approximation equations from Section 2. We find that these equations do give good approximations. In Section 5 we give an example of the use of the approximation equations in a decision problem, and in Section 6 we give a brief summary and conclusion.

2. STOCHASTIC APPROXIMATIONS TO THE KNOCK-ON EFFECT

Consider a single rail line or link between two stations (say stations A and B). There are two trains (numbered 1 and 2) travelling from station A to station B. The trains' departure times from station A are respectively $T_1$ and $T_2$, where $T_1 \leq T_2$, so that $H_1 = (T_1 - T_2)$ is the initial headway between the trains. If $T_1$ and $T_2$ are the scheduled departure times, it is assumed that the trains depart on time. The actual trip times (hence arrival times at station B) are random variables. Denote the trip time and arrival time of the 1st train by $t_1$ and $t_{1f}$, respectively, $(t_1 = T_1 + t_1)$. Similarly, let $t_2$ and $t_{2f}$ be the free running trip time and arrival time of the 2nd train $(t_2 = T_2 + t_2)$. That is, $t_2$ is the trip time that the 2nd train would incur if it was the only (or the first) train on the link (see the more precise definition in eqn (13) in Section 3). We assume stochastic independence of $t_1$ and $t_2$. However, due to knock-on effects, the probability distribution of the actual trip time and arrival time of the 2nd train is affected by those of the 1st train, and by the headway $H_1$. Denote the actual trip and arrival time of the 2nd train by $t_{2H}$ and $t_{2fH}$, the index $H_2$ indicating dependence on the headway. Of course, $t_{2H} = T_2 + t_{2H}$. Suppose for the moment that the speeds of trains 1 and 2 are constant along the link. Also, trains are not allowed to pass on the link. If train 2 catches up with train 1, which is ahead of it, it must reduce its speed to equal that of train 1 and, hence, arrive immediately after train 1, i.e., $t_{2H} = t_1$: this becomes $t_{2H} = (t_1 + h^2)$ if a minimum time headway
$h^a$ is required between the trains. On the other hand, if train 2 is not delayed by train 1 ahead, $t_{2,H_2} = t_2^*$. Combining the two cases we have (see Fig. 1),

$$t_{2,H_2}^a = \max\{t_2^*, t_1^* + h^a\}. \quad (1)$$

Subtracting $T_2 = (T_1 + H_2)$ from both sides, where $H_2$ is the headway on departure, gives the link trip time,

$$t_{2,H_2} = \max\{t_2, t_1 - H_2 + h^a\}. \quad (2)$$

However, in practice, train speeds are not constant along a link A to B. The scheduled speed varies with track gradient, curvature, etc., and is low as the train starts off from A or slows down to stop at B. There are also random variations in speed (random deviations from the scheduled speed), due to variation in operating conditions, weather, equipment, driver behaviour, etc. If the speed of the first train fluctuates, it may delay the following (second) train. This implies $t_{2,H_2}^a \geq t_2^*$; once $t_{2,H_2}^a$ exceeds $t_2^*$, it is often difficult or impossible to make up the gap again: trains in Britain are usually operating close to their maximum permitted line speed, hence recovering the lost time ($t_{2,H_2}^a - t_2^*$) would mean violating this maximum. After delaying the second train, the first train may then speed up again (or the second train may slow down) so that the first train may arrive well ahead of the second train (i.e., $t_{2,H_2}^a \geq [t_1^* + h^a]$). Combining the two cases we have,

$$t_{2,H_2}^a \geq \max\{t_2^*, t_1^* + h^a\}. \quad (3)$$

that is, (1) becomes an inequality rather than an equality.

The inequality time gap in eqn (3) is made larger by the fact that the headways enforced by train control systems (signals) are in fact usually distances between trains, rather than times. When a train slows down, this minimum distance takes longer to traverse, so that the resulting time headway between the trains increases. For example, if a train has to slow to half speed due to a train ahead, then the time headway between the two trains has to double to maintain the same distance headway. In this case, $h^a$ above is the required minimum time headway when the train is travelling at its normal (scheduled) speed (see 3-aspect signalling in Section 3.3).

Since eqn (3) sometimes holds as an equality and sometimes as an inequality, the r.h.s. is useful as a good lower bound on $t_{2,H_2}^a$. It may also be useful as an approximation to $t_{2,H_2}^a$ when there is little variation in train speeds. However, we are here interested in obtaining a better approximation to $t_{2,H_2}^a$, rather than only a lower bound. The basic hypothesis in this paper is that a good approximation to $t_{2,H_2}^a$ is obtained by simply replacing $h^a$ in eqn (1) with a larger shift or adjustment parameter $s > h^a$, so that it becomes,

$$t_{2,H_2}^a = \max\{t_2^*, t_1^* + s\}. \quad (4)$$

![fig1](image-url)
Subtracting \( T_2 = (T_1 + H_2) \) from both sides of eqn (4), where \( H_2 \) is the headway on departure, gives the hypothesis,

\[
t_z \cdot H_2 = \max \{ t_z, t_1 - H_2 + s \}. \tag{5}
\]

We define a good approximation in eqn (4) to be one for which the mean and variance of the error of approximation is small. Note that in eqn (4) we have implicitly included \( h' \) in the shift parameter \( s \). We are assuming that \( h' \) is a fixed parameter determined by signal spacings, etc. However, if \( h' \) is treated by operators as an adjustable control parameter, we can simply replace \( s \) with \((s' + h')\) in eqn (4), and the ensuing discussion and results are otherwise unchanged.

More specifically, we hypothesize that a fixed shift \( s \), independent of the departure headway, will give a good approximation (in the probabilistic sense) even when the trip times \( t_1 \) and \( t_2 \) are stochastic, and even when the train operating rules are as in the 2-aspect or 3-aspect signalling models set out in Sections 3.1 and 3.3 below. In the latter cases, \( t_1 \) and \( t_2 \) are in fact sums of the random variables representing trip times on the segments (track circuit segments in Britain) or pieces \( j = 1, \ldots, d \), of the link, and \( t_z \cdot H_2 \) also includes random wait times on any or all sections of the link due to train 1 ahead. The hypothesis that the shift \( s \) is independent of the departure headway is particularly important, since (a) it makes estimating a suitable value of \( s \) much more tractable, and (b) the exact departure headway is often not known in advance, being in itself a random variable.

Hypothesis (5) implies the hypotheses

\[
E[t_z \cdot H_2] = E[\max \{ t_z, t_1 - H_2 + s \}] \tag{6}
\]

and

\[
P_x[t_z \cdot H_2] = P_x[\max \{ t_z, t_1 - H_2 + s \}] \tag{7}
\]

where \( P_x[Y] \) denotes the \( x \)th percentile or quantile (e.g., the 90th percentile) of the probability distribution of \( Y \). Percentiles of trip time distributions are important in train planning and scheduling because a common reliability or performance rule is that the scheduled trip time be set equal to say the 90th or 95th percentile of the distribution of actual trip times.

These hypotheses raise the questions: [a] how accurate are such approximations, and [b] what is the best value for an adjustment constant \( s \) and how can it be determined? We concentrate on hypotheses (6) and (7) rather than (5), since these are more relevant and important in train planning and scheduling. One way to test the hypotheses is by collecting empirical data on trip times and departure headways and then using regression analysis to fit functions of the form (5)–(7), and computing measures of goodness of fit. This is described in Sections 2.1 and 4.

However, since we did not have suitable empirical data available, we used a Monte Carlo simulation approach to generate representative data. To do this, we simulated in detail the performances of a pair of trains under 2-aspect and 3-aspect signalling systems (Sections 3.1 and 3.3). We used this simulated data to generate graphs of \( E[t_z \cdot H_2] \) and \( P_x[t_z \cdot H_2] \), i.e., the left hand sides of eqns (6) and (7), against the departure headway \( H_2 \). We refer to these as the “true” relationships, of \( E[t_z \cdot H_2] \) and \( P_x[t_z \cdot H_2] \), to \( H_2 \). Then, using the same simulated data set, we applied regression analysis to estimate \( s \), computed the right hand sides of eqns (6) and (7), graphed these against \( H_2 \), and compared the results with the “true” relationships from the detailed simulations. From this we find that approximations of the form (5)–(7) do indeed yield good approximations (as shown later in Section 4).

The advantage of using eqns (6)–(7) to approximate \( E[t_z \cdot H_2] \) and \( P_x[t_z \cdot H_2] \) is that the r.h.s.’s of these can be computed or estimated by various means, if estimates of the probability distributions of the free running link trip times \( t_1 \) and \( t_2 \) are available. The
Effects of headways on knock-on delays of trains

r.h.s.'s of eqns (6) and (7) can be computed by numerical methods. If the distributions of \( t_i \) and \( t_2 \) happen to have simple functional forms, it may be possible to derive explicit functional forms for the r.h.s.'s of (6) and (7). For example, this can be done if the p.d.f.'s of \( t_i \) and \( t_2 \) are (or can be approximated by) shifted negative exponential distributions (i.e., fixed trip times plus a delay having a negative exponential distribution—see the example of Section 5). However, in the quite likely event that the p.d.f.'s of \( t_i \) and \( t_2 \) do not have a simple form, the convolution integrals needed to obtain the r.h.s.'s of eqns (6)–(7) may not have an analytic form. But, in this case numerical solutions (r.h.s.'s of eqns (6) and (7)) can be computed. Similarly, numerical solutions can be computed if the p.d.f.'s of \( t_i \) and \( t_2 \) are available only as general empirical distributions. This is likely to be the case, as detailed data on train link trip times is usually routinely collected for each train and each link.

In the remainder of this section we outline the way of choosing the adjustment constant \( s \) by regression that we later investigate in detail. An alternative heuristic method is briefly sketched in Section 3.2.

2.1. Estimating the adjustment constant \( s \) by regression

The parametric form of eqn (5) suggests using (nonlinear) regression to estimate the constant \( s \) in eqn (5). That is, rewrite eqn (5) as

\[
t_{2,H_2} = \max \{ t_2, \ t_i - H_2 + s \} + e_{H_2},
\]

where \( e_{H_2} \) is a random error. Let \((t_i^1, t_i^2), i = 1, \ldots, n\), denote a sample of \( n \) observed values of these random variables, and let \((t_{2,H_2}^i, e_{H_2}^i)\) be the corresponding values of the random variables \((t_{2,H_2}, e_{H_2})\), for a given headway \( H_2 \). That is,

\[
t_{2,H_2}^i = \max \{ t_2^i, \ t_i^2 - H_2 + s \} + e_{H_2}^i,
\]

for \( i = 1, \ldots, n \), and some relevant range of headways \( H_2 \in \mathcal{H} \). The sample size \( n \) may of course vary with \( H_2 \), but for simplicity we have not included this in the notation.

Then choose the value of \( s \) so as to minimize some function of the observed errors \( e_{H_2}^i \). That is, choose \( s \) so as to minimize

\[
\sum_{H_2 \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} f_{H_2}(e_{H_2}^i) = \sum_{H_2 \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} f_{H_2}(t_{2,H_2}^i - \max \{ t_2^i, t_i^1 - H_2 + s \})
\]

(9)

where say \( f_{H_2}(e_{H_2}^i) = (e_{H_2}^i)^2 \) or \(|e_{H_2}^i|\), or more generally \( f_{H_2}(\cdot) \) may include weights depending on headways \( H_2 \).

This regression approach requires that we have simultaneously observed values of \( t_i \), \( t_2 \), and \( t_{2,H_2} \). However, it is more likely in practice that we have separate samples for each of these random variables, i.e., a sample of free running trip times \( t_i \) and \( t_2 \), for trains which are not obstructed by trains in front of them, and a separate sample of trip times \( t_{2,H_2} \) for trains which are affected by trains in front. In this case we do not have observations \((t_i, t_2)\) corresponding to individual \( t_{2,H_2} \) observations, hence cannot compute \( e_{H_2}^i \) in eqn (9), hence cannot estimate \( s \) in this way.

However, the available data indicated above allow us to compute the mean or percentile of the right and left hand side of eqn (5), which suggests estimating the parameter \( s \) for eqn (6) or (7). And, as already noted in the introduction, eqns (6) and (7) are of more interest than eqn (5) in train planning. Following eqns (5) and (6) we can write,

\[
\bar{t}_{2,H_2} = \max \{ t_2, \ t_i - H_2 + s \} + \bar{e}_{H_2},
\]

(10)

for all \( H_2 \in \mathcal{H} \), where an overbar denotes the mean of a sample for a given fixed value of the headway \( H_2 \), e.g., \( t_{2,H_2} = \frac{1}{n} \sum_{i=1}^{n} t_{2,H_2}^i \). Then choose \( s \) so as to minimize some function of the observed mean errors \( \bar{e}_{H_2} \) in (10), that is, minimize
In Section 4 we use eqn (11) rather than eqn (9) to estimate \( s \), using data generated in a series of simulation experiments. We also use eqn (11), in Section 4, as a measure of goodness of fit of eqns (6) and (7) for \( s \) values obtained by methods other than regression.

The adjustment constant \( s \) may be different for each pair of train types on each link, hence a separate regression may have to be performed for each. We therefore also explore an alternative way of choosing the parameter \( s \) in eqn (5). We construct a relatively complex stochastic model of trains' interaction, work out some heuristics for choosing a value for \( s \), and test these heuristics in a series of pseudo-random computer experiments.

We follow this approach in Section 3.2.

3. TWO MODELS OF TRAIN INTERACTION

In this section we set up two stochastic models of the interaction of trains within a link. These models [a] provide a basis for simulation experiments (in Section 4) in which we test the simple formula (5) against such detailed models and [b] suggest a heuristic approach (in Section 3.2) to estimating the value of the adjustment constant \( s \) in eqn (5). Though in the remainder of the paper we explicitly discuss only the case of two trains, the two models hold good for any number of trains.

3.1. Two-aspect signalling model

In order to control the movement of trains within the link A \( \rightarrow \) B, and for safety reasons, the link is divided into a number of sections or signal blocks. The sections of the link are numbered from 1 to \( d \), section 1 being the closest to station A. There are \( k \) trains travelling from station A to B (trains 1, \( \cdots \), \( k \)). Their numbers indicate the order of departing from station A, i.e., train 1 departs first. Let \( T_i \) denote the scheduled departure time of the \( i \)th train, \( i = 1, \cdots, k \), and for \( i > 2 \) let \( H_i = T_i - T_{i-1} \) be the scheduled headway between the \( i \)th and \( (i - 1) \)th train. We may adopt the natural assumption that \( T_1 \leq T_2 \leq \cdots \leq T_k \) and consequently \( H_i \geq 0 \) for \( i = 2, \cdots, k \), although this is not necessary for any formal reason. As we are interested in studying the dependence of the knock-on effect on the initial headways, we assume that all trains are ready to depart exactly at their scheduled departure times.

The actual trip time of the \( i \)th train on the \( j \)th section, \( i = 1, \cdots, k, j = 1, \cdots, d \), is a random variable; let us denote it by \( t_{i,j} \). This random trip time does not include any waiting time within the \( j \)th section before the signal separating it from the \( (j + 1) \)th section. It reflects only factors such as the condition, topography, and length of the section; the train's technical characteristics; and driver behaviour. Denote by \( t_{i,0} \) \( i = 1, \cdots, k, j = 0, \cdots, d \), the actual trip time of the \( i \)th train, from its scheduled departure time from station A up to completing the \( j \)th section and entering the \( (j + 1) \)th one (or station B if \( j = d \)). Hence \( t_{i,0} \) is the time elapsed from the scheduled departure time of the train to the actual departure time from station A. The time \( t_{i,j} \) includes both this initial departure delay and the waiting time at the end of the \( j \)th section. Note that \( t_{i,0} \) is the total trip time of the \( i \)th train over the entire link. We shall also denote these total trip times by \( t_{i,H_{1}} \cdots H_{n} \) to emphasize the dependence on the initial headways \( H_{1}, \cdots, H_{n} \) as \( t_{i,H_{1}} \cdots H_{n} \) in Section 2. To make the analogy with Section 2 complete, and for convenience in describing figures, let \( t_{i,j} = T_{j} + t_{i,j} \).
Effects of headways on knock-on delays of trains

Possible time-space paths of two consecutive trains are shown on Figure 2. Horizontal segments of the paths represent waiting times of trains between sections. For instance, train $i$ waits at each signal block in the situation shown in the part (a) of Figure 2, while in part (b) it enters the $j$th section immediately.

The trip times of the 1st train are simply sums of the random variables $\tau_{i,j}$,

$$t_{1,0} = 0,$$
$$t_{1,j} = \sum_{r=1}^{j} \tau_{1,r}, \quad j = 1, \cdots, d. \quad (12)$$

We analogously define the free running trip time $t_i$ of any other train, that is, the time it would need to complete its journey in the same circumstances if it was the only (or the first) train on the link.

$$t_i = \sum_{r=1}^{d} \tau_{i,r}, \quad i = 1, \cdots, k. \quad (13)$$

Trip times $t_{i,j}$ for $i = 2, \cdots, k$, satisfy the following recursion:

$$t_{i,0} = \max\{0, t_{i-1,1} - H_{i}\}, \quad i = 2, \cdots, k,$$
$$t_{i,j} = \max\{t_{i,j-1} + \tau_{i,j}, t_{i-1,j+1} - H_{i}\}, \quad i = 2, \cdots, k, \quad j = 1, \cdots, d - 1,$$
$$t_{i,d} = t_{i,d-1} + \tau_{i,d}. \quad (14)$$

Recursion (14) is more straightforward when restated in terms of times $t_{i,j}^\theta$ (by adding $T_i$ to both sides of each equation in (14) and substituting $T_i - H_{i} = T_{i-1}$). For instance the first equation becomes

$$t_{i,0}^\theta = \max\{T_i, t_{i-1,1}^\theta\}.$$

This means simply that the $i$th train cannot depart from station A (because of a red signal) until the $(i - 1)$th one passes to the second section of the link.

It is also possible to express the total trip times $t_{i,d}$, $i = 2, \cdots, k$, explicitly in terms of random variables $\tau_{i,j}$ and headways $H_2, \cdots, H_k$. See eqn (15) for the case of $i = 2$.

3.2. Heuristics for the adjustment constant $s$

Our aim in this subsection is to draw simple conclusions from the description of the 2-aspect signalling model and use them to suggest a heuristic approach to the problem of estimating constant $s$ in eqn (5). We emphasize that it is not our purpose to derive the best possible heuristics for $s$, but only to convey the advantages of such an approach to
modelling trip times with knock-on effect. We did not wish to put too much effort into interpreting data from simulations rather than from reality.

Recursive equations (14) can be easily expanded to direct expressions for the partial trip times \( t_{i,j} \). For \( i = 2 \) using (12) and applying recursion (14) we have for instance:

\[
\begin{align*}
t_{2,0} &= \max\{0, \tau_{1,1} - H_2\} = \max\{0, \tau_{1,1} - H_2\}, \\
t_{2,1} &= \max\{t_{2,0} + \tau_{2,1}, t_{1,2} - H_2\} = \max\{\tau_{2,1}, \tau_{1,1} + \tau_{2,1} - H_2, \tau_{1,1} + \tau_{1,2} - H_2\}.
\end{align*}
\]

We omit the details of a simple induction procedure that gives the following equation for the total trip time \( t_{2,H_2} \):

\[
t_{2,d} = t_{2,H_2} = \max\{\tau_{2,1} + \cdots + \tau_{2,d}, \\
\tau_{1,1} + \tau_{2,1} + \tau_{2,2} + \cdots + \tau_{2,d} - H_2, \\
\tau_{1,1} + \tau_{1,2} + \tau_{2,2} + \cdots + \tau_{2,d} - H_2, \\
\cdots \\
\tau_{1,1} + \cdots + \tau_{1,m} + \tau_{2,m} + \cdots + \tau_{2,d} - H_2, \\
\cdots \\
\tau_{1,1} + \cdots + \tau_{1,d-1} + \tau_{2,d-1} + \tau_{2,d} - H_2, \\
\tau_{1,1} + \cdots + \tau_{1,d-1} + \tau_{1,d} + \tau_{2,d} - H_2\}.
\]

(15)

This expanded equation can provide us with a formal confirmation of intuitive inequality (3). Let \( H^* \) be the minimum value that can be assumed by \( \tau_{2,d} \), that is, the minimum arrival headway between the two trains. Skipping all arguments of the maximum in eqn (15) except of the first and the last one and using notation (13) we see at once that:

**Lemma 1**

\[
t_{2,H_2} \geq \max\{t_2, t_1 - H_2 + \tau_{2,d}\}.
\]

(See Figure 1 for interpretation.)

**Remark.** The well-known Jensen inequality states \( E[f(x)] \geq f(E[x]) \), where \( f \) is a convex function and \( x \) is a random variable or vector. Since "max" is a convex function, it might be supposed that one could take expectations through the above inequality and apply Jensen's inequality to the r.h.s. to obtain, \( E[\max\{t_2, t_1 - H_2 + \tau_{2,d}\}] \geq \max\{t_2, t_1 - H_2 + E[\tau_{2,d}]\} \) hence a lower bound on \( E[t_{2,H_2}] \). However, this is not valid, since \( \tau_{2,d} \) is a component of the random variable \( t \) so that the two arguments of max are not independent. If the latter inequality were true then, since its l.h.s. is a constant (an expected value) and its r.h.s. is a random variable, it would still be true if we replaced the r.h.s. with its expected value, thus, \( E[\max\{t_2, t_1 - H_2 + \tau_{2,d}\}] \geq E[\max\{t_2, t_1 - H_2 + E[\tau_{2,d}]\}] \). However, rather surprisingly, we can show that not only is this last inequality untrue, but it is true with the direction of the inequality reversed.

We consider two distinct general situations involving a pair of trains: case (a) the first train is the slower, or both trains have approximately the same speed, and case (b) the second train is the slower.

"Faster" and "slower" are rather imprecise terms, since trip times are stochastic. Nevertheless, we may safely assume that slower means "slower (or at least not faster) on each section of the link." In other words in, say, case (a) random variables \( \tau_{1,j} \) are in some sense stochastically greater than \( \tau_{2,j} \), e.g., for \( j = 1, \cdots, d \), and \( u \geq 0 \) we have \( P[\tau_{1,j} \geq u] \geq P[\tau_{2,j} \geq u] \) and consequently \( E[\tau_1] \geq E[\tau_2] \).

With this interpretation, in case (a) the last component of eqn (15), which equals \( (t_1 + \tau_{2,d} - H_2) \), tends to be greater than all but the first component, hence one may expect
Effects of headways on knock-on delays of trains

that \( t_{2,H_2} \) will be well approximated by eqn (5) with \( s \) close to \( E[T_{2,d}] \). However, even if the first train is the slower, the influence of the middle and later terms in eqn (15) may be not negligible, due to random variations. Especially the components close to the last one can have a significant effect on \( t_{2,H_2} \). The constant \( s \) should therefore exceed \( E[T_2,H_2] \). Some preliminary numerical experiments helped us to establish that \( s_0 = (E[T_{2,d}] + E[T_{2,d-1}]) \) gave a much better approximation in eqn (5). Hence, we may define the approximation \( \tilde{t}_{2,H_2} \) of \( t_{2,H_2} \) as

\[
\tilde{t}_{2,H_2} = \max\{t_2, t_1 - H_2 + s_0\}. \tag{16}
\]

For example, if all \( d \) sections of the link have more or less similar length, \( s_0 \) becomes

\[
s_0 = 2E[t_2]/d, \tag{17}
\]

In case (b), i.e., if the second train is the slower, the earlier terms in eqn (15) tend to be greater than the later ones. Such terms can be approximated as follows:

\[
\tau_{1,1} + \cdots + \tau_{1,m} + \tau_{2,m} + \cdots + \tau_{2,d} = t_1 + \tau_{2,m} + (\tau_{2,m+1} + \cdots + \tau_{2,d}) - (\tau_{1,m+1} + \cdots + \tau_{1,d}) \\
\approx t_1 + \tau_{2,m} + E[\tau_{2,m+1} + \cdots + \tau_{2,d}] - E[\tau_{1,m+1} + \cdots + \tau_{1,d}] \\
\approx t_1 + \tau_{2,m} + \alpha E[t_2] - \alpha E[t_1] = t_1 + \tau_{2,m} + \alpha(E[t_2 - t_1]) \text{ for some } \alpha \in (0, 1).
\]

For example, if all sections of the link have roughly the same length, the parameter \( \alpha \) is approximately \((d - m)/d\). Then by the same argument as for case (a), the max of these earlier terms in eqn (15) could be approximated by \( t_1 + s_0 + \alpha E[t_2 - t_1] \) where \( s_0 = (E[T_{2,1}] + E[T_{2,2}]) \), the expected trip times on the first few segments of the link. Combining cases (a) and (b), a rough estimate for \( s \) for both cases could have the form

\[
s = s_0 + \alpha(E[t_2 - t_1])^+. \tag{18}
\]

Here \((x)^+\) denotes \(\max\{0,x\}\) so that the second component of eqn (18) is present only in case (b). With this adjustment constant, the definition (16) becomes

\[
\tilde{t}_{2,H_2} = \max\{t_2, t_1 - H_2 + s\}. \tag{19}
\]

Choosing an appropriate value for \( \alpha \) is difficult. We again sought an answer to this problem in preliminary experiments and found that \( \alpha = 1/3 \) yields close approximations to the ones obtained from the regression procedure described in Section 2.1. We do not claim that eqn (18) with \( \alpha = 1/3 \) is in any way the "best" heuristic for \( s \). We merely point out that there are alternatives to the regression approach that allow \( s \) to be estimated with less computational effort, though with less accuracy.

Simulation experiments described in Section 4 test the performance of eqns (18) and (19), eqn (5) with constant \( s \) estimated using nonlinear regression, and eqn (5) with \( s = 0 \) for comparison. We also compare these with the "true" results obtained from detailed simulations of the 2-aspect and 3-aspect signalling models described in Sections 3.1 and 3.3.

3.3. Three-aspect signalling model

Consider the following modification to the 2-aspect signalling model defined in Section 3.1. The movement of trains is controlled by red, amber, and green signals. A red signal indicates that there is a train in the next section ahead. An amber signal indicates that there is no train in the next section, but there is a train in the section after that. A green signal indicates that there is no train in the next two sections. A train cannot pass a red signal. It can pass an amber one, but must then slow down. When a train enters a section, the signal at the beginning of the section turns from green or amber to red and at
the same time the signal at the beginning of the previous section turns from red to amber.

Also, if a train is in an “amber” section then it must reduce speed to a fraction $c_a < 1$ (“a” for “amber”) of its normal speed. For example, in the simulation experiments in Section 4 we use $c_a = 0.6$, which is typical of the value of $c_a$ enforced by British Rail. Train $i$ would then need time $\tau_{i,j}/c_a$ to traverse the “amber” section $j$. It is possible to define an “amber” factor $c_{ij}$ for each train $i$ and section $j$ separately or even an alternative set of amber trip times with different probability distributions. However, we confine ourselves to the simpler case, as the rules used in practice appear to be quite simple.

Definitions (12) and (13) remain unchanged in the new model, but recursion (14) requires redefining as follows,

$$t_{i,0} = \max\{0, t_{i-1,1} - H_i\}, i = 2, \cdots, k,$$

$$t_{i,j} = \begin{cases} \max\{t_{i,j-1} + \frac{\tau_{i,j}}{c_a}t_{i-1,j+1} - H_i\} & \text{if } t_{i,j-1} \leq t_{i-1,j+1} - H_i, \\ t_{i,j-1} + \tau_{i,j} & \text{otherwise,} \end{cases}$$

$$i = 2, \cdots, k, j = 1, \cdots, d - 1. \quad (20)$$

$$t_{i,d} = t_{i,d-1} + \tau_{i,d}.$$  

4. SIMULATION EXPERIMENTS

In a series of experiments (Scenarios 1 through 18 below) we simulated a multisection rail line under 2-aspect and 3-aspect signalling and compared the “true” simulated trip times $t_{2,H}$, with stochastic approximations to these of the form (5)–(7). We tried and compared three different forms of the approximations (5)–(7), namely, [a] approximations (5)–(7) using adjustment constant $s = 0$, [b] using $s = \bar{s}$ defined by (18), and [c] using $s$ obtained by the regression method discussed in Section 2.1.

We chose $s$ in (c) so as to minimize the mean absolute error in the approximation (5) over a range of $N$ different headways, ranging from $H_2 = a$ to $H_2 = b$ in $(N - 1)$ equal steps. That is, we chose $s$ to minimize,

$$\text{err} = \frac{1}{N} \sum_{H_2 = a}^{b} \left| \overline{t_{2,H}} - \max\{t_{2}, t_1 - H_2 + s\} \right|,$$  

(21)

where the overbar denotes the mean of a random sample for each value of $H_2$.

A single experiment is designed as follows. For a given value of $s$ and given scenario (number of track sections $d$, number of trains $k = 2$, and probability distributions of the link section trip times $\tau_{i,j}$, $i = 1, 2, j = 1, \cdots, d$) we:

[1] generate a pseudo-random sample of the random variables $\tau_{i,j}$ from eqn (22) or eqn (23) below,
[2] calculate the 2-aspect and 3-aspect signalling trip times $t_{2,H}$ from eqns (14) and (20), respectively,
[3] calculate free running trip times $t_1 = \Sigma_{j=1}^{d} \tau_{1,j}$ and $t_2 = \Sigma_{j=1}^{d} \tau_{2,j},$
[4] repeat calculations [1]–[3] for a large number (see below) of random samples and use the results to calculate $t_{2,H}$, and $\max\{t_{2}, t_1 - H_2 + s\}$ in (21),
[5] repeat [4] for $H_2$ varying from $a = 1$ minute to $b = 12$ minutes in steps of 0.5 minutes, hence calculate (21), for given $s$.

We repeat [5] for various values of $s$ to find the value of $s$ that minimizes eqn (21).

To evaluate the results, we calculated various characteristics of trip times (expectations, percentiles or quantiles, standard deviations, histograms of probability density functions) and graphed these against headways (see below). We performed 25,000 iterations in step
Effects of headways on knock-on delays of trains

above in each experiment (each scenario). We performed such a large number of iterations only (a) because it was easy to do so, (b) because we wished to ensure very smooth plots in figures such as 3 and 4, and (c) because we wished to be able to refer here to the sample mean $t_{2,H_2}$ as the true population mean $E[t_{2,H_2}]$. Smaller numbers of iterations (sample sizes) should be enough for the regression to give reliable estimates of $s$, i.e., to find $s$ to minimize eqn (21). If observed rather than simulated trip time data were used it is likely that only smaller numbers of observations would be available.

4.1. Link section trip times

As noted above, since data was not available to us for trip times $\tau_{i,j}$ of trains over short sections of track, we conducted simulations using some well-known theoretical probability distributions for $\tau_{i,j}$. Here we report results based on two of these distributions. Similar results were obtained for shifted atom-exponential distributions, but are not reported here.

1. Random variables $\tau_{i,j}$ having shifted exponential distributions

$$P\{\tau_{i,j} \leq u\} = \begin{cases} 0 & \text{if } u < k_{i,j}, \\ 1 - e^{-\lambda_{i,j}(u-k_{i,j})} & \text{if } u \geq k_{i,j}, \end{cases}$$

which implies minimum trip time $k_{i,j}$ and mean excess trip time $1/\lambda_{i,j}$.

2. Random variables $\tau_{i,j}$ having uniform distributions

$$P\{\tau_{i,j} \leq u\} = \begin{cases} (u - k_{i,j})/l_{i,j} & \text{if } k_{i,j} < u \leq k_{i,j} + l_{i,j}, \\ 0 & \text{otherwise}, \end{cases}$$

which implies minimum trip time $k_{i,j}$, maximum excess trip time $l_{i,j}$, and mean excess trip time $l_{i,j}/2$.

We “estimated” parameters for the above section trip time distributions very roughly, as follows. There is some evidence (e.g., Transport Statistics User Group, 1990) that for intercity trains in Britain the standard deviation of the link trip time is very roughly 5 minutes times the square root of the journey length in hours. As we express all times in minutes, the above becomes

$$\text{Var}[\tau_i] = (5/12) \times [\text{journey length in minutes for train } i].$$

If we assume that the minimum trip time is a measure of journey length then

$$\text{Var}[\tau_i] = (5/12) \sum_{j=1}^{d} k_{i,j}.$$  

We therefore set the parameters $\lambda_{i,j}$ and $l_{i,j}$, in eqns (22) and (23), respectively, so that (25) is satisfied, and at the same time made these proportional to $k_{i,j}$. In this way each simulation model can be defined by specifying the type of probability distribution, the number of sections $d$, and a list of constants $k_{i,j}$.

4.2. Results of simulations

Below we list some of the scenarios for which we conducted the above simulation experiments. The results of the other cases are very similar.

Scenario 1. A pair of two “fast” trains. The section trip times $\tau_{i,j}$ have shifted exponential distributions (22), the number of sections is $d = 20$, and the minimum section trip times $k_{i,j}$ are 1 minute for each train and section.

Scenario 2. A “slow” and a “fast” train. Minimum section trip times are 1.2 minutes for the “slow” train and 1 minute for the “fast” train. Otherwise, as Scenario 1, i.e., shifted exponential distributions (22), $d = 20$ sections.
Scenario 3. A “fast” and a “slow” train, as in Scenario 2, but with the order of the trains reversed.

Scenarios 4–6. As Scenarios 1–3, respectively, but the link is longer, consisting of \( d = 30 \) sections.

Scenarios 7–9. Respectively, “fast-fast,” “slow-fast,” and “fast-slow” pairs of trains on a link divided into \( d = 20 \) sections of irregular length. Minimum section trip times as in Table 1, with shifted exponential distributions (22).

Scenarios 10–18. Uniform distributions of the section trip times \( \tau_{i,j} \) as in eqn (23). All other characteristics as in Scenarios 1–9, respectively.

In all cases we compute the error (21) over the range of headways from \( a = 1 \) minute to \( b = 12 \) minutes, with the step of 0.5 minutes. This is because [a] headways less than 1 minute are very unlikely in practice, and [b] as is illustrated by Figure 3, for headways of over 12 minutes the knock on effects (and errors in approximating these) in scenarios presented here are negligible.

The results of the simulation experiments are summarized in Table 2 and Figures 3 and 4, with all times in minutes and seconds. These indicate the goodness of fit of our three different ways of approximating trip times with knock-on effects. Column 1 in Table 2 contains the number identifying the scenario and the two-letter code indicating the relative train speeds (e.g., “ff” = “fast–fast”). Column 2 gives the mean free running trip time \( \mathbb{E}[t] \) for the second train—this helps put the other figures in perspective. Columns 3 and 4 show the results of approximating the trip time, \( t_{2,H_2} \), with formula (5) with the adjustment constant \( s = 0 \). The label “err2” denotes the value of the average error (21) with \( t_{2,H_2} \) obtained from the 2-aspect signalling model (rounded to whole seconds). “Err3” has the same meaning for the 3-aspect signalling model. The next three double columns summarize the results for the approximations based on adjustment constant \( s \) chosen [a] heuristically as in (18), [b] by nonlinear regression as in steps [4]–[6] above for the 2-aspect signalling model, and [c] by nonlinear regression for the 3-aspect model. The corresponding values of the average error function (21) is again in columns labelled “err2” and “err3”.

It can be seen from the last three double columns in Table 2 (and from Figures 3 and 4) that errors err2 and err3 are relatively small. That is, the regression method yields adjustment constants \( s \) that, in turn, yield good approximations to the “true” link trip time \( t_{2,H_2} \), including knock-on delays: the “true” trip time being given by the detailed simulation model. The heuristic method (18) yields a much smaller improvement in the approximations to \( t_{2,H_2} \). To put the errors in perspective, in the regression case they are all much less than one percent of the link free-running trip times in Column 2. Also, these errors are almost all 5 to 10 times smaller than the errors that would be obtained if we had not introduced the adjustment constant \( s \) (setting \( s = 0 \) as in the first double column).

It can be seen from Figure 3 that the performance of both approximation methods is not as good for very short headways as for medium and longer headways. The performance of the regression method for small headways can be further improved by giving more weight to such headways in (21). However, in practice, headways smaller than a few minutes are usually not allowed, hence are not very relevant.

To further illustrate the improvement of approximation provided by the adjustment constant \( s \), we compare (in Table 3) the “true” and approximated expected trip times of

<table>
<thead>
<tr>
<th>Train type</th>
<th>Minimum trip times ( k_{ij} ), ( i = 1, \ldots, 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>&quot;Fast&quot;</td>
<td>0.5</td>
</tr>
<tr>
<td>&quot;Slow&quot;</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Effects of headways on knock-on delays of trains

- True trip time $t_{2,H_1}$ in the 2-aspect signalling model, see Section 3.1.
- True trip time $t_{2,H_3}$ in the 3-aspect signalling model, see Section 3.3.
- Approximation to $t_{z,H_1}$ from (5) with $s = 0$.
- Approximation to $t_{z,H_3}$ from (19), with $s$ defined by (18).
- Approximation to $t_{z,y}$ from (5), with $s$ obtained by regression, to minimize (21).

The two lines are for 2- and 3-aspect signalling models, respectively.

Free running trip time $t_z$ of the second train (not dependent on headway), see definition (13).

Range of summation in (21)

Fig. 3. Expectations of the trip times and their approximations for the 2nd train in “fast-fast” Scenario 1, with 2-aspect signalling.

The 2nd train in the six “fast–fast” pairs of trains. The approximation is given by the r.h.s. of (6), with $s$ obtained by the regression method and the headway $H_z$ fixed at 3 minutes. To make comparisons easier, all trip times are scaled by dividing by the corresponding free running trip times. Note that the columns labelled “$E[t_{2,H_3}]$” contain the “true” expected trip times for 2-aspect and 3-aspect signalling respectively. Again it can be seen that (6) gives a good approximation to the true $E[t_{2,H_3}]$ obtained from the detailed simulations. Also, it can be seen that the approximation error ($E[t_{2,H_3}] - $approx. (6)) is small compared to the knock-on delay ($E[t] - E[t_z]$).

Table 2. Goodness of approximations to “true” simulated trip times, in mins:secs.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$E[t_z]$</th>
<th>$s = 0$</th>
<th>Heuristic (18)</th>
<th>Reg'n, 2-asp.</th>
<th>Reg'n, 3-asp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ff</td>
<td>32:44</td>
<td>31:12</td>
<td>3:17 :06</td>
<td>2:56 :05</td>
<td>5:31 :12</td>
</tr>
<tr>
<td>2 sf</td>
<td>32:44</td>
<td>31:12</td>
<td>3:17 :06</td>
<td>2:56 :05</td>
<td>5:31 :12</td>
</tr>
<tr>
<td>3 fs</td>
<td>37:56</td>
<td>34:49</td>
<td>5:33 :05</td>
<td>4:50 :04</td>
<td>8:19 :14</td>
</tr>
<tr>
<td>7 ff</td>
<td>58:21</td>
<td>51:08</td>
<td>2:31 :04</td>
<td>2:30 :04</td>
<td>5:04 :10</td>
</tr>
<tr>
<td>8 sf</td>
<td>28:21</td>
<td>24:04</td>
<td>2:31 :04</td>
<td>2:30 :04</td>
<td>5:04 :10</td>
</tr>
<tr>
<td>10 ff</td>
<td>42:22</td>
<td>42:02</td>
<td>3:14 :12</td>
<td>3:36 :07</td>
<td>7:02 :15</td>
</tr>
<tr>
<td>11 sf</td>
<td>42:22</td>
<td>42:02</td>
<td>3:14 :12</td>
<td>3:36 :07</td>
<td>7:02 :15</td>
</tr>
<tr>
<td>18 fs</td>
<td>40:51</td>
<td>38:13</td>
<td>5:19 :07</td>
<td>4:34 :05</td>
<td>8:39 :16</td>
</tr>
</tbody>
</table>

Based on $s = 0$, $s$ from (18), and $s$ from regression analysis.
Perhaps the best illustration of the various approximations is Figure 4. It presents densities of trip times and their approximations for the 2nd train in Scenario 1. The headway is fixed at 3 minutes. For clarity, only the results for 3-aspect simulation model are given, but results are similar for 2-aspect simulation. The figure suggests that approximations of the form (5) with appropriately chosen constant s can serve as good estimators of position parameters (such as expectation, mode, and percentiles) of the probability distribution of \( t_{f,H_f} \).

More figures illustrating the goodness of fit of the approximations can be found in Appendices of Carey and Kwiecinski (1992). The figures there illustrate and compare the "true" expected trip times and our estimated approximations to these for various scenarios. They show that even though the constant s minimizes the error (21) based on the mean values of the samples, the same value of s can also be used to obtain good estimates of, for example, percentiles (quantiles) of the trip times. Of course, if it is quantiles that are of most interest, then the approximation could be further improved by using quantiles rather than mean values in the error (21) to be minimized in the regression process.

Equation (5) can also be used in computing approximations to spread or dispersion parameters of the trip time, such as the standard deviation. However, it is not better than other approximations for this purpose. We note however that the standard deviation is not much affected by the headways in the simulations we conducted. For example, for the second train in Scenario 1, the "free running" standard deviation is about 3 minutes and the standard deviations computed via the various approximations are all within a 20-second band around this.

5. EXAMPLE APPLICATION TO A SIMPLE DECISION PROBLEM

Various uses of the approximations developed in this paper are indicated in the introduction, and in (i)-(iv) in the conclusion. An example of one of these uses is as follows—this is of type (i) from the conclusion Section 6.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Free run. ( E[t_f] )</th>
<th>Approx. (6) with ( s = 0 )</th>
<th>Regression for 2-asp.</th>
<th>Regression for 3-asp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>101.7</td>
<td>106.0</td>
<td>104.9</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>101.7</td>
<td>104.6</td>
<td>104.4</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>101.6</td>
<td>104.6</td>
<td>104.9</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>101.3</td>
<td>104.9</td>
<td>104.7</td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td>101.3</td>
<td>104.3</td>
<td>104.0</td>
</tr>
<tr>
<td>16</td>
<td>100</td>
<td>101.2</td>
<td>104.6</td>
<td>104.4</td>
</tr>
</tbody>
</table>
Consider the following problem. Two trains are scheduled to travel from station A to station B using the same line, and cannot pass each other en route. The second train is ready to depart on time, but the first train will not be ready to depart until exactly $L$ minutes late. If the delay $L$ is sufficiently large it may be that the aggregate cost of lateness for both trains at B can be reduced by letting the second train depart first. We wish to calculate the lateness $L = L_0$ at which it becomes just profitable to swap from the scheduled departure order. If there is no uncertainty concerning link trip times, then the calculation of $L_0$ is simple arithmetic, assuming that the cost per minute of lateness for each train is known.

If the trip times are stochastic, as in Section 3, we can calculate $L_0$ so as to minimize expected costs. To do this we can use the stochastic approximations to link trip times given in this paper. We here use approximation (17)–(19) to determine $L_0$ so as to minimize expected costs for a particular numerical example.

As before, let $T_i^d$ and $T_i^a$, respectively, be the scheduled departure time from station A and scheduled arrival time at station B for train $i = 1, 2$, and assume $T_1^d < T_2^d, T_1^a < T_2^a$. Let $t_i$ be the actual free running trip time of train $i$, and set $F_i(u) = P\{t_i \leq u\}$. Let $\bar{t}_{i,H}$ be an approximation to the actual trip time of train $i$, including knock-on delays, when it departs second, and $H$ is the actual departure headway between the trains. From (19),

$$\bar{t}_{1,H} = \max\{t_1, t_2 - H + s_{11}\},$$
$$\bar{t}_{2,H} = \max\{t_2, t_1 - H + s_{12}\},$$

$s_{11}$ and $s_{12}$ being appropriate adjustment constants. There is a required minimum headway between the departures of the trains, and let this be $h_{12}$ if train 1 departs first and be $h_{21}$ if train 2 departs first.

Let the cost to be minimized be the expected lateness of both trains at station B. For each train,

$$[\text{lateness at st. B}] = ([\text{dep. time}] + [\text{trip time}] - [\text{sched. arrival time}])^+, \quad (26)$$

where $(x)^+ = \max\{0, x\}$. Let the expected value of the delay (26) for train $i$ be $C_i^{(\text{swap})}(L)$ if the trains are swapped, and be $C_i^{(\text{noswap})}(L)$ if they are not. The expected total costs associated with swapping or not swapping are thus,

$$C^{(\text{no swap})} = w_1 C_1^{(\text{no swap})} + w_2 C_2^{(\text{no swap})},$$
$$C^{(\text{swap})} = w_1 C_1^{(\text{swap})} + w_2 C_2^{(\text{swap})}, \quad (27)$$

where $w_1$ and $w_2$ are cost weights. These weights may represent the average number of passengers on each train or some other measure of train importance or priority. Formal expressions for these expected costs are derived in Carey and Kwiecinski (1992) and to save space are not repeated here.

The least cost decision rule for swapping trains is: for given delay $L$ of train 1 swap trains 1 and 2 if and only if the total cost $C^{(\text{swap})}$ is less than $C^{(\text{noswap})}$. We wish to determine the range of delays $L$ for which the trains should be swapped. The problem may be impossible to solve analytically, but one can easily determine whether to swap the trains by using simple numerical procedures, for any given parameters $T_i^d, T_i^a, T_i^e, T_i^s, h_{12}, h_{21}, s_{11}$, and $s_{12}$, and probability distributions ($F_i$ and $F_j$) of $t_i$ and $t_j$.

We solved the above problem (determined $L_0$) for the following specific example. Let

<table>
<thead>
<tr>
<th>Table 4. The constants used in the example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>i = 1 (j = 2)</td>
</tr>
<tr>
<td>i = 2 (j = 1)</td>
</tr>
</tbody>
</table>

TR(B) 28:4-B
the free running trip times $t_i$ of both trains have shifted atom-exponential distributions as follows.

$$
P\{t_i \leq u\} = \begin{cases} 
0 & \text{if } u < k_i, \\
(1 - (1 - p_i)e^{-\lambda_i(u-k_i)}) & \text{if } u \geq k_i,
\end{cases}$$

which implies a minimum trip time $k_i$ occurring with probability $p_i$, and a mean excess trip time $(1 - p_i)/\lambda_i$. Such distributions may occur if trains are not allowed to arrive earlier than a certain time $k_i$, so that there is a fraction $p_i$ of trains arriving on time and an exponential tail of later arrivals. The values of the constants in our numerical example are listed in Table 4. We choose these values to ensure "realistic" trip time and reliability values as in Table 5.

To relate Table 4 to Table 5 recall that for the atom-exponential distribution $E[t_i] = (1 - p_i)/\lambda_i + k_i$ and $\text{Var}[t_i] = (1 - p_i)/\lambda_i^2$ (see (24)). The constants in Table 4 ensure that train 1 is faster than train 2 and that $\text{Var}[t_i] = \frac{1}{2}\lambda_i k_i$ for both trains. They also imply the reliability figures for the free running case given in the last two columns of Table 5. The headway constants $h_i$ in Table 4 are set equal to the expected trip time of train $i$ on the first section of the link, assuming that the link is divided into 30 sections of equal length. The constants $s_{ij}$ are calculated from (17)-(18).

Since the train scheduled first is faster than the second one, it is natural to associate a relatively greater cost with lateness of the first train. Thus, let the cost ratio be $w_1/w_2 = 5$ in (27), which is quite likely in train operations. Then it can be shown that the numerical solution is $L_0 = 16 : 56$ (Carey and Kwiecinski, 1992). In contrast, if the adjustment constants $s_{ij}$ are left out (set equal to 0), then the "swap" point turns out to be $13 : 13$, which is $3 : 43$ less. As a second example, let $w_1/w_2 = 10$ in (27), which again is not unusual. Then $L_0 = 27 : 05$ if we again compute it using the $\delta$ from (17)-(18), and $L_0 = 14 : 57$ is we compute it ignoring $s$ (i.e., letting $s = 0$). Note that in this case the value of $L_0$ is almost 100% higher when taking account of the adjustment constant $s$, which is a substantial difference. If we further increase the ratio $w_1/w_2$, then the effect of $s$ on the optimal swap point $L_0$ and the loss caused by an incorrect decision (swapping or not swapping) also increase.

Thus train 2 should wait for the departure of train 1 if train 1 is no more than 17 minutes late, otherwise the trains should be swapped. The scheduled headway is 5 minutes, hence delay of the first train may delay the departure of the second train by up to 12 minutes, under the optimal policy. This delay in swapping the trains is because the first train is much faster than the second. If the first train is only a small amount $L$ late, then letting the (slower) second train go in front of it would substantially delay it, without bringing much benefit to the second train.

6. SUMMARY AND CONCLUSIONS

This paper deals with the problem of approximating the link trip time of a train when it is preceded by another train which can delay it. The trains are subject to 2-aspect or 3-aspect (or more aspects) signalling on each section of the link, are not allowed to pass each other, and their trip times on each section of the link are subject to random variation. We found that the link trip time $(t_{2,H})$, including knock-on delays caused by a preceding train, can be well approximated by a simple formula,

$$\tilde{t}_{2,H} = \max\{t_2, t_1 - H_2 + s\},$$

Table 5. Statistics for the trip times in the example

<table>
<thead>
<tr>
<th>$i$</th>
<th>$E[t_i]$</th>
<th>$\text{Var}[t_i]$</th>
<th>$\text{St. Dev.}[t_i]$</th>
<th>Late arr.</th>
<th>Late $\geq 5$ min</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32.5</td>
<td>12.5</td>
<td>3.54</td>
<td>18%</td>
<td>5%</td>
</tr>
<tr>
<td>2</td>
<td>39</td>
<td>15</td>
<td>3.81</td>
<td>17%</td>
<td>5%</td>
</tr>
</tbody>
</table>
where \( t_1 \) and \( t_2 \) are the free running link trip times of the trains, \( H_2 \) is the departure headway between the two trains, and \( s \) is an adjustment constant. The approximation can be applied sequentially to a series of trains on a link.

We discuss methods of determining the constant \( s \) (e.g., by nonlinear regression and by heuristic arguments) and test these and the resulting trip-time approximations \( \tilde{t}_{L,H_2} \) in a series of computer simulations. The regression method minimizes the approximation error, and is therefore the most accurate method. It also allows one to use cost or error functions to emphasize certain subranges of headways for which accuracy may be more important. However, the regression method requires that samples of link trip times for pairs of trains are available for a range of different headways. The heuristic approach is less accurate, but is simpler and requires only data that is more easily available (such as mean trip times, etc.) for a given headway. The approximations can be used in a number of ways, for example (in no particular order):

1. To formulate and solve simple train planning, dispatching, or control problems involving a pair of trains, taking account of knock-on delays varying with the departure headway. An example of such an application is given in Section 5.
2. To simplify simulation models by reducing or removing the need for detailed modelling and simulation of train movements, signalling system, etc., on each subsection of each link. By simplifying the treatment of each link, stochastic simulation of rail networks is greatly simplified, and the amount of data required is greatly reduced.
3. To generate deterministic approximations to expected knock-on delays. These approximations may be used to improve headway allowances in train network planning models, which are usually deterministic.
4. To simplify estimating the capacity (throughput) of a link.

The last two uses of the approximation are being further explored. Further work that would also be of interest includes, (a) investigating the effects of more complex signalling and control systems on the approximation, (b) developing improved heuristics for the adjustment constant \( s \), (c) estimating \( s \) and testing the approximations using observed train data rather than simulation generated data.

Acknowledgements—This research was supported by U.K. Science and Engineering Research Council (SERC) grants GR/F/91407 and GR/H/50432, and by cooperation with British Rail. The authors would like to thank the SERC and British Rail for their support. An earlier version of this paper was presented (but not published) at the 24th Annual Universities Transport Studies Group (UTSG) Conference, held at the University of Newcastle-upon-Tyne, England, 5-8 January 1992. The authors would also like to thank the anonymous referees for their helpful comments and suggestions.

REFERENCES

Crainic T. G. and Querin L. (1987) Approximation of over-the-line delay for single line, single track rail transportation. Publication No. 519, Centre de Recherche sur les Transports, University of Montreal, Montreal, Canada.