Forecasting delays on railway sections

T. Huisman

Department of Innovation, Railned B. V., Utrecht, The Netherlands.

Abstract

This paper presents a stochastic model to forecast delays on a section in a railway network. For this model, a simple algorithm is derived to calculate delay distributions, both for transient and stationary (equilibrium) behaviour. In an application to a railway section of the Dutch railway network, we show how the algorithm can help to improve the quality of the timetable.

1 Introduction

Delays play an important role in the quality of service that a railway company offers to its passengers. In literature, therefore various models have been proposed to forecast these delays: network-wide models, like the simulation model of [4], or the max-plus models [3], and models to analyse specific components of the network, such as the queueing model of [2] for railway sections. This paper extends the results of [2] allowing analysis of the evolution of delays in time (transient behaviour), and providing more efficient calculation of delay distributions for cyclic timetables, i.e. timetables that consist of a fixed, repeated pattern.

The paper is organised as follows. Section 2 presents our model of a railway section. In section 3, we derive an algorithm to compute delay distributions, which is applied to a railway section of the Dutch railway network. This application shows that, for a cyclic timetable, delay distributions fairly rapidly converge to a fixed distribution: they do not change in time any more, and thus a stationary situation is reached. Section 4 provides an algorithm to compute these stationary delay distributions directly for a cyclic timetable. We apply this algorithm to the same railway section, and show how this algorithm may help us to improve the planning. Section 5,
finally, discusses some extensions of the model.

2 Model and definitions

We consider a unidirectional track of a railway section connecting two distinct locations \( X \) and \( Y \). The track may include some stations, as long as overtaking at this stations is impossible, or not allowed (Figure 1).

![Figure 1: A unidirectional track from \( X \) to \( Y \) with two intermediate stations.](image)

Trains using this track are labelled 1,2,... in order of their departure from \( X \). For train \( n = 1,2,... \), we define:

- \( D^{\text{plan}}_n \) as its planned departure time from \( X \),
- \( A^{\text{plan}}_n \) as its planned arrival time at \( Y \),
- \( R^{\text{plan}}_n \) as \( A^{\text{plan}}_n - D^{\text{plan}}_n \), i.e. its planned running time,
- \( R^{\text{min}}_n \) as its minimal running time, i.e. the minimal time at which train \( n \) can theoretically run from \( X \) to \( Y \),
- \( H_n \) as the minimal headway time between train \( n \) and train \( n+1 \), i.e. the minimal time that should separate train \( n+1 \) from train \( n \) throughout the section due to safety regulations.

We will assume that the planning is such that it can - at least, theoretically - be realised without delays, i.e. that the difference between two consecutive planned arrival or departure epochs exceeds or equals the minimal headway time, and that the planned running time exceeds or equals the minimal running time.

In practice, however, the planning will not always realised without delays. Due to, for example, weather conditions or prolonged dwell times, actual running times may exceed the planned running time, resulting in trains arriving at \( Y \) later than their planned arrival times. These stochastic phenomena are described by the following random variables for train \( n = 1,2,... \).

\( S_n \) denotes the disturbance of train \( n \), defined as the prolongation of train \( n \)'s minimal running time, that is not due to other trains, but to external causes.
\(L_n^{\text{prim}}\) denotes the primary delay of train \(n\), i.e. the prolongation of train \(n\)'s planned running time due to its disturbance. 

\(D_n^{\text{act}}\) is defined as train \(n\)'s actual departure time from \(X\). 

\(A_n^{\text{act}}\) is defined as train \(n\)'s actual arrival time at \(Y\).

The relation between the actual characteristics of the trains and their planning is assumed to be as follows. First, we assume that trains depart from \(X\) on time, i.e. \(D_n^{\text{act}} = D_n^{\text{plan}}\). Furthermore, we assume that

\[
L_n^{\text{prim}} = \max(0, S_n - (R_n^{\text{plan}} - R_n^{\text{min}})), \quad n = 1, 2, \ldots; \tag{1}
\]

the primary delay thus is that part of the disturbance that exceeds the difference between planned and minimal running time. This difference is often referred to as the slack time. Finally, we assume that

\[
A_1^{\text{act}} = D_1^{\text{plan}} + R_1^{\text{plan}} + L_1^{\text{prim}}, \tag{2}
\]

\[
A_n^{\text{act}} = \max \left( L_n^{\text{plan}} + R_n^{\text{plan}} + L_n^{\text{prim}}, A_{n-1}^{\text{act}} + H_{n-1} \right), \quad n \geq 2. \tag{3}
\]

Observing that the primary delay is defined as the prolongation of its planned running time due to its disturbance, relation (2) is immediate by our assumption that trains depart from \(X\) on time. For train \(n = 2, 3, \ldots\) this relation is extended by 3: these trains can not only be delayed by their disturbance, but also by previous trains. If the latter is the case, a train slows down, adopts the speed of its predecessor, and follows it by a minimal headway time. The resulting delay is often called secondary delay.

The assumptions made above allow us to derive a recursive formula for the delays, presented in the following theorem. This theorem can be proven in analogy with Theorem 2.1. of [2].

**Theorem 2.1** For \(n = 1, 2, \ldots\), let \(L_n\) denote the delay of train \(n\) upon arrival at \(Y\), i.e. \(L_n = A_n^{\text{act}} - A_n^{\text{plan}}\). Then \(L_n\) satisfies the following recursive formulation.

\[
L_1 = L_1^{\text{prim}}, \tag{4}
\]

\[
L_n = \max \left( L_n^{\text{prim}}, L_{n-1} - B_{n-1} + R_{n-1}^{\text{plan}} - R_n^{\text{plan}} \right), \quad n = 2, 3, \ldots. \tag{5}
\]

where \(B_n\) is defined as \(D_{n+1}^{\text{plan}} - D_n^{\text{plan}} - H_n\), the planned buffer time between train \(n\) and train \(n + 1\).

**Remark 2.1** Theorem 2.1 relates the delays in out track model to the sojourn time in an infinite server resequencing queue with service times \(L_n^{\text{prim}}\) and interarrival times \(B_{n-1} - R_{n-1}^{\text{plan}} + R_n^{\text{plan}}\). Note that these interarrival times are non-negative (as interarrival times in a queuing model are supposed to be) by the assumption that the planning can be theoretically realised without delays.
3 Results for general planning

In this section, we assume that the planning of all trains using the railway section is given. Under this assumption, the recursive relation that we derived for the delays in Theorem 2.1, can be translated into a corresponding relation for the delay distributions.

**Theorem 3.1** Assume that the disturbances are independent random variables with distribution \( s_n(t) \) for train \( n \), i.e. \( P(S_n \leq t) = s_n(t) \). Then the delay distribution \( \ell_1(t) \) is given by \( s_1(t + (R_n^{\text{plan}} - R_n^{\text{min}})) \), and the delay distribution \( \ell_n(t) \) of train \( n = 2, 3, \ldots \) can be computed from \( \ell_{n-1}(t) \) by

\[
\ell_n(t) = s_n(t + (R_n^{\text{plan}} - R_n^{\text{min}}))\ell_{n-1}(t + B_{n-1} - R_{n-1}^{\text{plan}} + R_n^{\text{plan}}).
\]

**Proof.** The proof follows from Theorem 2.1 by conditioning on the disturbances, and relation 1 between primary delays and disturbances.

The results of Theorem 3.1 allow evaluation of the evolution of the delays during a day. We illustrate this with an application to a railway section in the Dutch railway network: the section between Utrecht and Geldermalsen. This is a 25 kilometer section containing three intermediate stations (Figure 2).

![Figure 2: The railway section Utrecht-Geldermalsen.](image)

The planning of this section consists of a cyclic pattern of four trains, that is repeated every 30 minutes throughout the day. Thus, train 5 has the same properties as train 1, train 6 the same as train 2, etc. Trains \( n \) and train \( n + 4 \) are therefore said to belong to the same train class. We thus have a cyclic timetable with four train classes, whose planning characteristics are provided in Table 1. All times in this table are in minutes.

We consider the first five hours of a day (from 6:00 to 11:00) with the following scenario for the disturbances. During the first hour all trains experience a mean disturbance of one minute. During the morning rush hours (7:00 to 9:00) the mean disturbances of intercity trains remain one minute; regional trains, however, then have larger disturbances due to prolonged dwell times at the intermediate stations: for class 2 trains, the mean disturbance is two minutes, for class 4 trains, having an additionally stop in...
Lunetten, it is three minutes. After the morning rush hours, the mean disturbances are as before: one minute for all trains.

For the distribution of the disturbances a truncated geometrical distribution on \([0, 0.1, 0.2, \ldots, 30]\) minutes has been chosen: the probability on a disturbance then is inversely proportional to the size of the disturbance. We have chosen a discrete distribution to simplify calculations.

Theorem 3.1 now provides the results of Figure 3, where the probabilities are depicted that the delay of train \(n = 1\) to 40 does not exceed three minutes. This performance measure - to which we will refer shortly as the punctuality - has been chosen, since the Dutch railways consider a train on time, when its delay is smaller than three minutes.

![Figure 3: Punctuality of trains \(n = 1\) to 40.](image)

We observe that the punctuality of the 8 trains in the first hour is very high; this is due to the relatively small disturbances. In particular the class

<table>
<thead>
<tr>
<th>Class</th>
<th>Service</th>
<th>Dwell stations</th>
<th>(R_{c}^{\text{min}})</th>
<th>(R_{c}^{\text{plan}})</th>
<th>(B_{c})</th>
<th>(H_{c})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercity</td>
<td>-</td>
<td>14.2</td>
<td>15</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Regional</td>
<td>Houten, Culemborg</td>
<td>18.7</td>
<td>20</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Intercity</td>
<td>-</td>
<td>13.8</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Regional</td>
<td>Lunetten, Houten, Culemborg</td>
<td>20.4</td>
<td>24</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Train classes for the section Utrecht-Geldermalsen.
4 regional trains hardly suffer any delay: the small disturbances, in combination with a relatively large slack time implies that primary delays of class 4 trains are small. Moreover, the regional trains of class 4 are the slowest trains on the track; they will therefore not often be delayed by other trains, i.e. also the secondary delays are small. Mathematically, this is explained by the recursive formulation (5) of the delays: the combination of a small disturbance and a large planned running time results in small delays.

When the morning rush hours start, however, the increased disturbances of regional trains cause a rapid decrease of punctuality. This increase of disturbances not only influences the regional trains: also intercity trains are less punctual, since they are more often delayed by regional trains. After approximately half an hour (i.e., one cycle), the punctuality of each train class remains more or less constant, and some differences between trains with similar service, but a different class appear: intercity trains of class 1 are less punctual than intercity trains of class 2, and similarly regional trains of class 2 are less punctual than regional trains of class 4. The latter is easily explained by the larger slack time of class 4 trains; the difference between the two intercity classes may be caused partly by similar reasons, but the main influence can be found by a closer look at the planning (Table 1): from the planned running times and the buffertimes, we can derive that the class 1 trains are planned immediately after class 4 train at the end of the track, whereas class 3 trains have a few minutes after class 2 trains. Class 1 trains thus have a large probability on secondary delays.

After the morning rush hours, the punctuality rapidly increases, and, after approximately one cycle of half an hour, returns to the same level as before the rush hours.

Figure 3 provides another interesting observation: when the disturbance distribution of each train class remains constant, after approximately half an hour also the resulting delay distribution of each train class remains constant: a steady state situation appears to be reached. We will adress this topic in the next section.

### 4 Results for cyclic planning

We have seen that, when the timetable is cyclic, delay distributions appear to converge fairly rapidly to a fixed limiting distribution. This observation can be theoretically explained by the relation of our track model with the infinite server resequencing queue (Remark 2.1); using the convergence result for this type of queueing model that is derived in [1], it can be shown that, for a cyclic timetable, the delay distribution of each train class converges to a fixed distribution when time tends to infinity, under the condition that at least one of the planned buffertimes is strictly positive. In the remainder of this paper, we will assume that this condition is satisfied.

The correspondence of the delays in our model with sojourn times in the infinite server resequencing queue also allows us to obtain the stationary
distributions of each train class directly. Again using the results of [1], the following theorem can be derived.

**Theorem 4.1** Assume that the planning is cyclic with \( C \) train classes \( 1, \ldots, C \), assume that, for \( c = 1, \ldots, C \), the distribution of the disturbance of a class \( c \) train is given by \( s_c(t) \), and that disturbances are finite, say bounded by a value \( T \). Moreover, assume that \( B_C - R^\text{plan}_C + R^\text{plan}_1 > 0 \). Then the limiting distribution \( \ell_c(t) \) of a class \( c \) train exists, and these distributions can be obtained recursively, backwards in \( t \), and for each \( t \) in the order \( c = 1, \ldots, C \) by

\[
\ell_n(t) = s_n(t + (R^\text{plan}_n - R^\text{min}_n))\ell_{n-1}(t + B_{n-1} - R^\text{plan}_{n-1} + R^\text{plan}_n)
\]

otherwise. Here, the index \( c \) should be read modulo \( C \).

**Remark 4.1** The assumption that disturbances are finite is clearly no limitation in practice. The assumption that \( B_C - R^\text{plan}_C + R^\text{plan}_1 > 0 \) may seem somewhat cryptic, but is in practice always satisfied: one can show that, by a suitably relabelling (shifting) of the trainclasses, this condition holds if at least one of the buffertimes is strictly positive. Note that this assumption is needed to allow for a recursive computation of (6).

Theorem 4.1 allows an efficient, direct computation of the limiting distributions of the delays, thus enabling a simple evaluation of the punctuality of a railway section, and a quick comparison of different plannings. Let us illustrate this by returning to the railway section Utrecht-Geldermalsen, and see if the punctuality of this section can be improved by some small modifications in the planning of Table 1.

We first evaluate the punctuality of the current planning using the rush hour scenario for the disturbances, i.e. mean 1, 2, 1, and 3 minutes for the respective train classes. The second column of Table 2 provides the results; note that these numbers correspond with the punctuality of trains 11 to 24 in Figure 3: in our previous example, the limiting regime thus was reached indeed.

<table>
<thead>
<tr>
<th>Train class</th>
<th>Punctuality</th>
<th>Punctuality after modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.87</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td>3</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>0.89</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 2: Delay probabilities and secondary delay probabilities for the 2000/2001 timetable
As mentioned in the previous example, the main source of delays for the two regional train classes are primary delays, not secondary delays; improving punctuality of these trains thus can be realised most efficiently by enlarging their slack times, so that they can compensate more of the disturbances. The two intercity train classes, however, have relatively small primary delays; their main source of delays thus is given by secondary delays. This holds especially for the class 1 intercity trains, which are planned at minimal headway time after the class 4 regional train at arrival in Geldermalsen.

Motivated by these considerations, we now modify the planning as follows. The planned running times of class 2 and class 4 are enlarged by one minute; the planned buffer time between class 3 and class 4 is reduced by two minutes, and the planned buffer time between class 4 and class 1 is increased by two minutes. Note that after this modifications, the timetable still is conflict free.

The third column of Table 2 shows that these modifications have the desired effect: the punctuality has been improved significantly.

5 Concluding remarks

We have provided a model for a railway section, that allows simple calculation of delay distributions, and we have shown how this model can help to reduce the propagation of delays on a railway section. These results, however, are based on some simplifying modelling assumptions; especially the assumption that trains enter the section on time may in practice not always realistic. We are currently developing an extension of the model, in which this assumption is relaxed. Another extension that we are working on is the modelling of a siding along the track, at which train may overtake.

References


