PROBABILISTIC PREDICTION OF CUMULATED DELAYS INDUCED BY TECHNICAL FAILURES IN URBAN RAIL NETWORKS

PREVISION PROBABILISTE DES DELAIS CUMULES DANS LES RESEAUX FERRES URBAINS LIES A DES DEFAILLANCES TECHNIQUES

René Valenzuela et Pierre Dersin
Alstom Transport Information Solutions
48 Rue Albert Dhalenne, 93400 Saint-Ouen, France

Résumé

On présente ici un modèle probabiliste de prévision du retard total cumulé induit par les défaillances techniques dans un réseau ferroviaire urbain. Cette prévision est donnée par l’estimation de la distribution de probabilité du retard total cumulé. L’estimateur est calculé à partir d’un échantillon de retards générés par des simulations de Monte Carlo d’un modèle de propagation de retards. On utilise ici cette estimation pour construire un intervalle de confiance autour de la valeur moyenne, mais on pourrait tout aussi bien l’appliquer également à des questions d’analyses de risques. L’accent est mis sur les défaillances techniques du système de signalisation, mais la méthodologie est applicable aux défaillances d’autres éléments du système de transport, tels que, par exemple, le matériel roulant.

Summary

In the communication a probabilistic prediction of the total cumulated delay induced by technical failures in urban rail networks is presented. The prediction is based on the estimated probability distribution of the total cumulated delay. The estimator is calculated from a sample of delays generated by the Monte Carlo simulation of a propagation delay model. In this work, the estimation is used to provide an interval estimate of the expected value but it can also be used to answer risk analysis based questions. The focus is on technical failures associated with a signaling system but the methodology is equally applicable to other types of failure modes associated with the equipment of an urban rail network (rolling stock).

Introduction

In recent years calls, for tenders for new urban projects have included increasingly often system requirements related to the quality of service the system will be able to provide. One of these new requirements is expressed as a bound on the total cumulative delay experienced by the system in a predefined interval of time. The total cumulative delay is defined as the sum of the differences between the actual mission end time and the nominal mission end time (i.e. a trip that ends its mission at 12:02 when its end mission time according to the nominal timetable was 12:00 has incurred a 2 [min] delay) for all the trips taking place within the considered time interval.

Accurate prediction of this indicator is both important and complex. For the infrastructure provider it is important because contractual penalties can be enforced if the specific requirement is not satisfied, for the operator of a rail network it is important for both robust timetable design (D’ariano et al 2009) and delay management (Cacchiani et al 2012). It is also complex because it depends on the technical characteristics of the whole urban system (which define the different service affecting failure modes with their corresponding failure rates), the characteristics of the line where the system operates and the characteristics of nominal operation (timetable) and degraded operation.

In spite of this need, currently available commercial tools are built with the objective of analyzing or optimizing the nominal behavior of the system (for timetable design or capacity analysis for instance) and as such are not intrinsically suited to handle the inclusion of failures and/or their effects (degraded modes of operation). Moreover, in order to perform their analysis most of the tools require detailed information of the system (track gradients, position of signaling elements along the track, etc.) which might not be available at the tender stage of a project.

With this in mind, in this work a custom simulation model is built which requires less detailed information about the system to be created and which uses Monte Carlo simulation to obtain a bound of the effects due to not modeled uncertainties.

The paper is organized as follows; first an overall description of the system being modeled is provided. Then, a review of three different approaches to model railway operations is presented. After that, an overall view of the modeling framework used in this work is shown. Finally, an example of application of the framework is described.
System Description

In the context of main lines operations, railways are typically operated according to a master timetable. This timetable represents a conflict-free coordination of train paths. These timetables have already embedded some slack time to manage train delays. The objective of having these buffer times between train paths is to prevent that slightly delayed trains immediately hinder following trains. Experience has shown that even though overall line capacity decreases by including these slack times the added overall timetable robustness to unforeseen events makes it desirable. However, depending on the distribution of running time supplements and buffer times, a single delayed train may cause a domino effect of secondary delays over the entire network due to train connections and route conflicts (Goverde 2010).

Although in principle urban metro operations also rely on a timetable during their normal operation there are several differences in their behavior when faced to an unforeseen event (failure or other). One such difference is the typical headway between consecutive trains. In urban metro this headway can be in the order of the couple of minutes during peak hours (which occur for several hours a day). Another difference is the recovery actions taken after an event has occurred. In urban operations if the timetable has been strongly perturbed then an operation based on maintaining a fixed headway (service regularity) might be more reasonable.

Modeling Railway Operations

Three different approaches to modeling railway operations are discussed: max-plus algebra, synchronous simulation of hybrid dynamics models and asynchronous simulation of discrete event models.

One approach that has been studied to model delays in railway networks is the max-plus algebra approach. The approach is based on the fact that a scheduled railway system can be modelled efficiently as a discrete-event dynamic system using max-plus algebra (Heidergott et al 2005). Furthermore, in the case of periodic timetables (like the ones usually implemented in railway operations) the dynamics of train sequences and synchronizations are simple recursive linear equations in max-plus algebra (Goverde 2010).

The deterministic max-plus model assumes that each event waits until all required resources are available and maintains the fixed order of all events according to the given timetable. Furthermore, late trains use available recovery times and propagation of delays is reduced by available buffer times. The max-plus recursions can be used to effectively compute the propagation of train delays over the railway network in both time and space (Goverde 2010).

This modeling methodology offers the advantage of providing a comprehensive set of analytical tools (Heidergott et al 2005) that allow both the easy calculation of performance metrics and at the same time the ability to better understand the behavior of a particular system or the possible behaviors of the general class of systems (eigenvalue analysis). One disadvantage of the framework is that does not allow for the easy introduction of more complex behaviors like: handling of priorities for conflict resolution, trains overpassing each other, provisional service (for urban operations).

A second alternative to railway operation modeling is to use synchronous models. These models capture the dynamics of the system by time-stepping through it. From the point of view of the implementation they follow an objects-oriented or agent-based approach. Commercial tools that use this type of models include RailSim TPC and OpenTrack. From the point of view of simulation time this modeling framework is not so efficient because it the time increments at each step have to be small to maintain the accuracy of the results. On the other hand, very detailed models can be obtained if enough effort is consumed during the modeling stage.

A third alternative to railway operations modeling is to use asynchronous simulation. This methodology emulates the timetabling process in which the trips of the timetable are immediately represented in the time-distance plane and then are modifying accordingly so as to achieve feasible timetables. In this modeling paradigm the presence of a failure can be represented on an extra temporary constraint imposed either on the trackside or in the rolling-stock dynamics which then modify the trips affected by it.

In this work the asynchronous simulation modeling paradigm is used. In our implementation the main element of the system is a trip. A trip is defined by the tuple \((R, P, L)\), where \(R\) is a graph representing the rail network, \(P\) is the representation of the path followed during the trip and \(L\) represents the locomotive unit that performs the trip. It must be noted that in this representation a trip path is only aware of the presence of stations or other signalling equipment by the constraints on velocity that they impose on its path.
The path is represented by the concatenation of three different types of segments: a constant-speed segment, a constant-acceleration segment and a waiting time segment. This representation simplifies the generation of the paths but does not allow for the easy modeling of trips where trains do not move at constant speed or acceleration.

In the graph modeling the network, vertices are timetabling points and edges are a single track between the two vertices. In the more general case, the appropriate representation should be an undirected multigraph where edges have an identity. The weight associated to each edge is the distance between the nodes it connects.

In the following section we proceed to describe how the failure modes are represented in the framework.

### Modeling of Failure Modes and Failures

Two different failure mode models are included: one model to represent failure of onboard equipment and the other one to represent failure of trackside equipment. The distinction is made to make easier the handling of the associated failure events.

The elements that characterize the failure mode models are:

1. A probabilistic model of the time to failure. Currently only the exponential model is included.
2. A model to represent the duration of the failure (or time to restoration).
   - For trackside equipment any probabilistic model can be used. A fixed value (deterministic) is also implemented. It is possible to parameterize the deterministic model by the time of the day. This parameterization is useful when modeling restoration actions that occur for instance only at the end of the day.
   - For onboard equipment, a probabilistic model is used to describe the restoration time once the maintenance personnel is onboard and able to perform the actions.
3. A model of the degraded operational procedure that needs to be applied to the affected trips.

### Modeling Effect of Failures: Degraded Operational Procedures

From the point of view of a trip, the occurrence of a failure is an event that may result in the partial or complete modification of its path. Assuming that a failure has already taken place then the implementation of the effect of the failure is a three step process:

1. Identify the set of trips that is directly affected by the failure. For instance, in the case of rolling stock failures, the directly affected trip is simply the trip associated to the rolling stock element in fault. On the other hand, in the case of the failure of trackside equipment (e.g. point machine, signals) the directly affected trips are all the trips that pass through the portion of the line where the element is fault is located (this could be in a single sense or in both of them).
2. Apply the degraded operational procedure to the directly affected trips. In the current implementation, this degraded operational procedure is a function of the failure mode associated to the failure and possibly on the ranking of the specific trip with respect to the set of directly affected trips (i.e. the degraded operational procedure for the first trip that encounters the failure may be different to the degraded operational procedure implemented by the second tip encountering the failure).
3. Modify the path of the trips indirectly affected by the failure event. In particular, there may be some trips that need to be delayed because the trips ahead of them have been directly or indirectly affected (delayed) by a failure that has occurred further ahead on the line. This modification include the possibility of trip cancellation if the operational procedures of the network allow for this possibility.

To identify the set of trip directly affected by the failure in the case of trackside equipment the procedure is simply to select all the trips whose path passes through the affected section of the track during the time on which the failure is considered to be in effect.

The modification of the path due to the application of the degraded operational procedure is a little bit more involved and will be described in the following subsection.

1. **Trip Modification Procedure**

The modification procedure is based on the fact that at the trip level, every degraded operational procedure can be expressed in terms of a concatenation of the following actions:

1. Modify (reduce/increase) current speed to a given value with a given constant acceleration
2. Continue in the current state a certain amount of time or through a given distance

For instance a degraded operational procedure that calls for the trip to:
1. Apply an emergency break
2. Have the driver perform a set of diagnostic actions and communicate them to the control center
3. Restart movement on a different operation mode (on which the train unit is moving slower)
4. After a signal on the track has been crossed return to normal operation

can be represented by a concatenation of the following operations: reduce speed to zero with emergency break acceleration, wait X minutes (can be deterministic or probabilistic), increase speed to a certain value at the nominal trip acceleration, wait until the head/tail of the train unit crosses the signal, increase speed to the nominal speed at that section of the line.

However, since failures can occur randomly along the track it could happen that in some portion of the trip that would be affected by the application of the degraded operational procedure there are constraints of the nominal path which are more stringent. For instance, in the previous example of a degraded operational procedure it could occur that the nominal path is reducing its speed because there is a station coming nearby while at the same time the degraded operational procedure calls for an increase in speed after the stoppage due to the failure.

In the simulation framework developed during this work, the final path is chosen as follows: look at the velocity as a function of the distance in the nominal case and superimpose the velocity profile imposed by the degraded procedure at the appropriate location along the line. The velocity profile of the modified trip is simply the velocity profile where the velocity is the minimum value between the velocity imposed by the nominal case and the velocity imposed by the degraded operational procedure. An example follows to illustrate the methodology.

![Figure 1: A nominal trip](image)

Figure 1 shows a trip path in three different dimensions: time, distance, and velocity. The trip has a constant velocity section followed up by a constant acceleration section, a waiting time section, another constant acceleration section and then another constant velocity section. It is clear how the concatenation of such segments can represent a trip with multiple stops at different stations.
Figure 2 illustrates in purple the representation of the operational degraded procedure that needs to be put in place after the occurrence of a failure at the location denoted by the star on the figure. Note how blindly applying the procedure to the nominal trip would result in the trip skipping the station at the location $S_2$ because due to the location of the failure the degraded operation procedure calls for the trip to be returning to normal operation at that portion of the track.

Figure 3 shows the modified trip that result from selecting the minimum value of the velocity at each section of the line. It is clear that the modified trips respects all the given constraints (both of the nominal trip and the one imposed by the degraded operational procedure due to the presence of the failure).

The modification of the trips indirectly affected by the failure is made following an analogous procedure but looking at constraints in the time-distance plane instead of in the distance-velocity plane.

**The Simulation Algorithm**

To perform the calculation of the metrics of interest a Monte Carlo simulation approach is followed. The use of Monte Carlo is motivated by the idea that a failure of the same type can produce considerably different total cumulated delays depending both on the position along the track and the time of the day at which it occurs.

The delay prediction model uses, as inputs, information from the FMECA, line layout, line operating characteristics, line degraded modes procedures and kinematic characteristics of the rolling stock operated on the line.
With this information, the delay model produces an estimation of the total cumulative delay due to a failure of a specific failure mode, occurring at a given time of the day in a specific position along the track and when the line is operating under a given timetable.

A Monte Carlo simulation engine is then used to generate a random sequence of failure instances and to collect the sampled cumulative delays. The empirical cumulative distribution function is used as an estimate of the true cumulative distribution. The overall calculation algorithm works as follows,

1. Define number of replications $N_{REPL}$
2. For each replication:
   a. Sample total number of failures and failure mode types
   b. For each failure mode sample a failure
   c. Sample a timetable (for the case where there are different timetables in place at different portions of the year)
   d. Apply failure to timetable
   e. Calculate cumulative delay $D_i$

**Results**

To study the capabilities of the framework the custom built delay model is applied to a main line system. Two types of trains circulate along the track, we will refer to them as either slow trains or fast trains. Besides the difference in top speed between the two types of trains there is also a difference in the maximum acceleration of which they are capable of achieving. Some of the trains stop at two different stations, others only in one station and finally there are some that do not stop along the route. Figure 4 shows a small portion of the nominal timetable. The colors in the figure identify both the direction of the trip and whether the locomotive unit performing it is of the slow or fast variety.

![Nominal timetable (portion)](image)

The calculation is based on failures due to 40 possible failure modes. These failure modes represent a typical set of service affecting failure modes.

The effect of a random failure on the trip paths and its associated delay is shown in Figure 5 (a). The results of a Monte Carlo simulation of this system are shown in Figure 5 (b).
Application of the modeling framework confirms the existence of variability on the total cumulative delay due to the heterogeneity in effective headways between consecutive trains of a typical timetable. When comparing between different failure modes, the effect of the restoration time on the total cumulative delay can also be quantified.

**Conclusion**

In this work an asynchronous modeling framework for the evaluation of delay in a railway network was introduced. This modeling framework includes an approach to handle conflict resolution and degraded operational procedures that need to be implemented after failures occur in the network. Evaluation of the metrics is performed through Monte Carlo simulation. The probabilistic approach is used so that the inherent random nature of failures can be taken into account but also because this randomness plus local effects of the network topology and timetable put in place result in variability in the performance metrics.

This variability in performance metrics was visualized in the calculation of the total cumulative delay metric. The possibility of estimating this distribution is of great importance to manufacturer’s and operators alike because it allows them to estimate possible penalties due to delays in normal operation.

It was found as well that the methodology based on trip representation by concatenation of segments is a feasible alternative to a the representation of rolling-stock dynamics (differential equations) moving in time. A disadvantage of the methodology is that it becomes increasingly harder to implement when trying to represent globally applied procedures or decentralized decision making for the operations of the rail network.

**References**