Impact of calibration of perturbations in simulation: the case of robustness evaluation at a station

Marie Milliet de Faverges, Christophe Picouleau, Giorgio Russolillo, Boubekeur Merabet, Bertrand Houzel

Impact of calibration of perturbations in simulation: the case of robustness evaluation at a station

Marie Milliet de Faverges [a,b], Christophe Picouleau [a], Giorgio Russolillo [a]
Boubekeur Merabet [b] Bertrand Houzel [b]
[a] CEDRIC laboratory, CNAM Paris, France
[b] DGEX Solutions, SNCF Réseau, Saint-Denis, France

Abstract
This paper deals with robustness evaluation at station, and in particular for the train platforming problem (TPP). This problem consists in a platform and route assignment in station for each scheduled train. A classical robustness evaluation is simulation: simulated delays are injected on arriving and departing trains then propagated, and results are averaged on a large number of trials. A robust solution of the TPP aims to limit the total amount of secondary delays. However, a simulation framework at station is difficult to calibrate: it requires a realistic delays generator and an accurate operating rules modeling.
This paper proposes an original simulation framework using classical statistical learning algorithms and calibration assessment methods to model simulation inputs. This methodology is applied on delay data to simulate delay propagation at station. It highlights the importance of delay calibration by showing that even slight miscalibration of inputs can lead to strong deviations in propagation results.

Keywords
Simulation, platforming problem, Calibration, Machine learning, Delay Distribution

1 Introduction

Robustness evaluation is a central topic for both academical and industrial actors in the railway field. Resources are saturated, demand is increasing and the network is congested, while investments are rare and expensive. This leads to strong pressure on infrastructure manager and railways companies to respond to these new problems. The challenge is particularly important at main stations: they form bottlenecks on the railway network, and delays propagate fast due among others to shared infrastructure, rolling stock planning and passenger activity. It is crucial to optimize railway operations robustness at station to limit the impact of perturbations.

The recent availability of delay data is a promising opportunity for that. Delays are recorded at different points of the railway network, allowing to have a better comprehension and analysis of perturbations occurrences and propagation. This is useful to improve railway models accuracy at different levels (delay distributions, operating rules,..) or to imagine new strategies based on these records.

This paper presents preliminary results on possible utilization of Machine learning approaches for robustness evaluation at station. It proposes a simulation framework using classical statistical learning algorithms and calibration assessment methods to model simu-

lation inputs. The learning model estimates individual probabilities of delay of each train based on the context, and the quality of the predicted probabilities is assessed independently of the simulation. These predictions are then used to simulate delay propagation at station. The machine learning approach is compared with other delay models. This experiment highlights the importance of calibration by showing that even a slight miscalibration of inputs can lead to strong deviation in propagation results.

This study is structured as follow: section 2 presents a short overview of existing works on railway simulation for robustness evaluation, section 3 describes the case study and the chosen methodology. Delay modeling work is shown in section 4 and delay propagation algorithm in section 5. Experiments are conducted in section 6 and results are discussed in section 7.

2 Related Work

This research proposes a new way of assessing the calibration of the perturbations generator in a simulation framework. Reviews of related studies conducted on both simulation for railway robustness evaluation and delay modeling are provided in this section.

2.1 Simulation for robustness evaluation:

A robust solution of an operations research problem is in general defined as a solution that will remain feasible when input parameters experience small variations. In railway research, schedules are usually not feasible anymore when disturbances occur, and robustness is more about finding a solution that can be recovered with limited use of dispatching (delay propagation, rescheduling, reordering, etc). In particular for railway station operations, a robust solution generally aims to reduce delay propagation and the amount of secondary delays (Caprara et al. 2010; Armstrong and Preston 2017).

There are two main ways to evaluate robustness of schedules, and in particular at station. The first one is to define reliability indicators based on characteristics of the schedule (headways, residual capacity, margins, etc). For instance Carey 1999 proposes deterministic reliability measures based on headways spreading in station. Performance indicators are easy to compute, but only give a partial vision of the robustness as they do not reflect traffic performances. The second one is simulation. It requires extensive description of the infrastructure, operating rules and perturbations distribution, but gives a more realistic and global evaluation of the ability of the solution to deal with small perturbations in real conditions.

It is however important to calibrate the parameters of the simulation tool correctly, especially the operating rules and the disturbances distribution used for sampling (Koutsopoulos and Wang 2007). Setting operating rules is complicated: real-time dispatching decisions are various (reordering, rerouting, event cancellation, etc) and it may be difficult to anticipate agents' choice in real-time. Moreover, these dispatching actions are not compatible with robustness evaluation concepts (reduced delay propagation with limited use of delay management). It must be decided during the simulation tool design what are the available decisions, and in which conditions they are applied. Carey and Carville 2000 use simulation and delay propagation algorithm to analyze reliability of routing and platforming solutions. Two operating frameworks are studied: one with fixed platform assignment and the other one with the possibility of platform changes, reducing strongly the amount of knock-on delays. On the other side, calibration of the disturbance distribution is also a key

element: simulation aims to estimate the behavior of the solution in real operations. For that, simulated delays must be reasonable, otherwise results will not be relevant. Usually, perturbations are generated according to a given probability distribution and then applied on the solution. For instance, Carey and Carville 2000 generate small delays using a uniform distribution and a beta distribution and apply them to randomly chosen train at each step of the simulation.

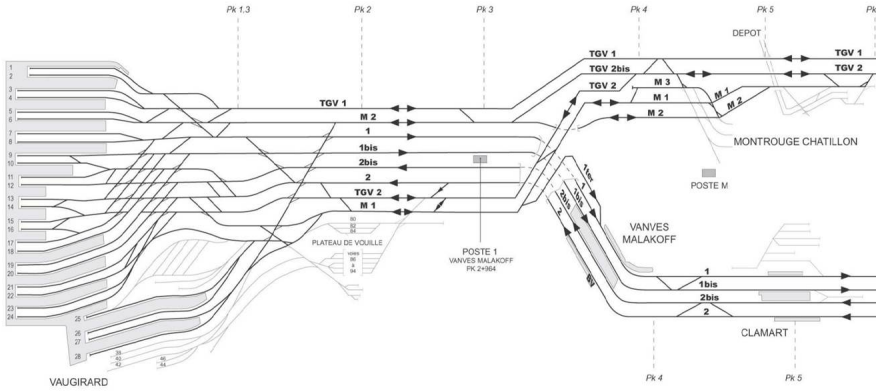
For the past few years, railway data, and in particular historical records of realized circulations have been more available. This is a promising opportunity for reliability measurement, and in particular for the sampling in simulation tools. Indeed, it is difficult to generate by hand a distribution that is concordant with reality, and actual observations of the network may help to find a way to generate reasonable d elays. To deal with this issue, Landex and O. A. Nielsen 2006 calibrate the delay distribution and rules of operations by comparing actual outputs and simulated outputs. They repeat this step several times before using their module to evaluate the robustness of timetables. Büker and Seybold 2012 express the issue of unknown primary delay distribution, since primary and secondary delays are not separated in operational data. Similarly, they compare key performance indicators based on results of simulations with operations records in order to calibrate the distributions. Koutsopoulos and Wang 2007 propose a calibration methodology based on minimization of the error between observed and simulated measurement. Larsen et al. 2014 model dwell times with a Weibull distribution for robustness evaluation using simulation. The location and shape parameters are estimated by maximum likelihood for peak hours and off-peak hours using records of arrival and departure times. Cui, Martin, and Zhao 2016 present an original method using reinforcement learning to automatically calibrate initial delays of simulation tools. Disturbances parameters are updated until convergence of a cost function. They present an application on a real network, where two parameter (mean delay and probability of delay) are tuned per combination type of train/ type of disturbance.

2.2 Delay Modeling

Train delay modeling is a well studied subject in railway research. Many studies have focused on finding an adequate distribution for empirical observations of train delays. Goverde 2005 uses a Kolmogorov-Smirnov test to assess the goodness-of-fit of state-of-the-art distributions (normal or negative exponential) on different types of delays recorded at the Eindhoven station (arrival delays, arrival non-negative delays, departure delays and dwell time excedents). Yuan 2006 evaluates different candidate distributions for delay records from the station the Hague, with one test per train type and direction. The Weibull, Gamma and log-normal distributions fit non-negative arrival and departure delay data well based on the Kolmogorov-Smirnov test. Briggs and Beck 2007 model delays in UK with q-exponential laws. Bergström and Krüger 2012 compute maximum likelihood estimation of the coefficients of lognormal, negative exponential and power-law distributions, and compare them graphically with observations from the Swedish railway network. Wen et al. 2017 show that primary delay durations are better fitted with a log-normal distribution than a Weibull one, even for data from different stations or during different period of the day. Harrod, Pournaras, and B. F. Nielsen 2018 show that delays on the Danish network are beter modeled with mixed distributions of lognormals than with a negative exponential distribution.

However results may depend on a large number of factors, like the type of delay (arrival, departure, dwell time), the range of values, location (station, line, etc), type of train,

Figure 1: Montparnasse station layout



operating rules, etc and may not be transposable from one case study to the other. For high-speed arrival non-negative delay data from the Montparnasse station, Faverges et al. 2018a compare state-of-the art distributions based on the Akaike Information criterion (AIC), and choose the negative binomial and the lognormal distributions to model delays.

3 Problem description

3.1 The platforming problem

The train platforming problem consists in routing trains through station and affecting them platforms. First solutions must be given months before operations, but adjustments can be done until a few days in advance. This problem is known to be NP-complete (Kroon, Romeijn, and Zwaneveld 1997). Finding solutions can be very challenging for main stations due to traffic density and a complex i nfrastructure. The train timetable is given, so arrival and departure time are fixed and solutions must satisfy commercial, security, resources and passenger flow c onstraints. This problem has been well studied with various approaches, for instance with MILP (Mixed Integer Linear Programming) formulation, constraint propagation or greedy heuristic (Sels et al. 2014).

SNCF Réseau, the french infrastructure manager, has recently developed a tool, OpenGOV, to solve the route and platform assignment problem at station. It is based on an extensive description of the station layout (platforms, paths, conflicts between resources) and the description of the different constraints. The problem is solved using MILP. Binary variables match trains with incoming path, platform and outgoing path. Two conflicting resources (crossing paths, tracks, platform, etc) cannot be affected to trains in the same time window whose size depends on the type of trains and the type of resources in conflict.

The case study is the Montparnasse station in Paris, France. This is a terminal station with 28 platforms and about 500 incoming and outgoing paths. There are about 700 sched-

uled trains per day, with suburban trains, high-speed trains and intercity trains. The models of infrastructure and operational constraints implemented in OpenGOV for the Montparnasse station are used in this study. Moreover, only passenger trains are considered for initial delay distribution and delay propagation. Indeed, other trains are more flexible and have a lower priority. They often experience variations in travel time (positive and negative delays) to adapt to other trains. Therefore, observation data of technical trains are unusable and dispatching rules are too complex to be modeled.

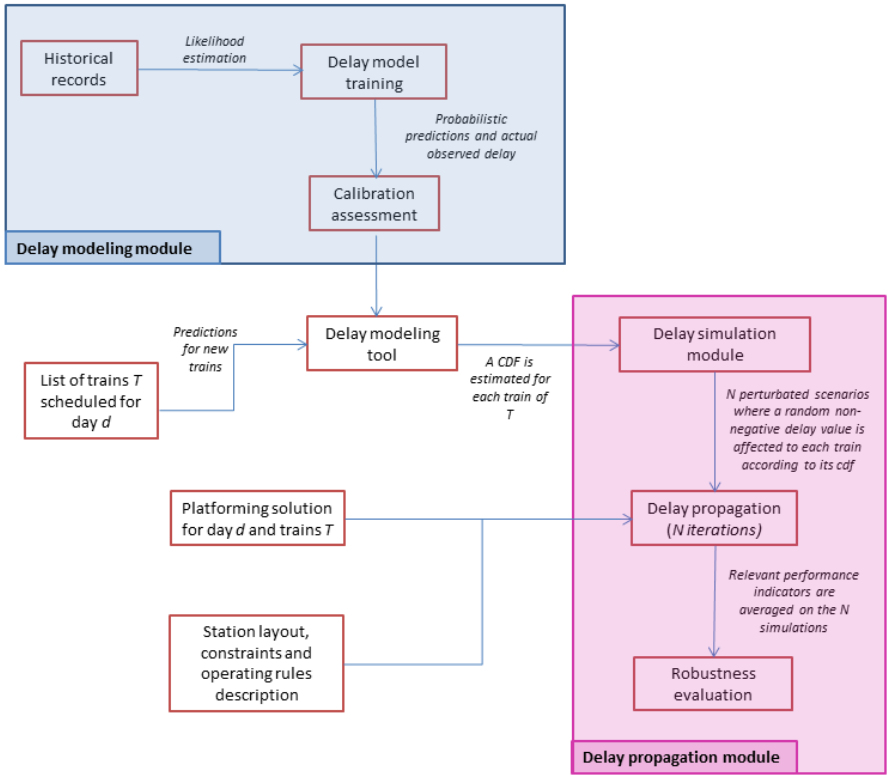
3.2 Proposed simulation methodology

This work takes advantage of the large amount of data collected on the network to build a calibrated and data-driven simulation framework. The methodology is summarized on the figure 2 . Historical records at a station are used to built a probabilistic model for initial delays. This model is then applied to simulate new delay samples (perturbations scenarios) for trains of a given day. N scenarios containing one delay value (non negative and often equal to zero) for each train of the day are obtained. For each of these scenarios, delays are propagated according to a platforming solution and given operating rules. The performance of the solution is then evaluated by averaging results of the delay propagation on the N iterations.

In this approach, the model for perturbations simulation is studied independently of the operating rules modeling, and in particular its adequacy is assessed before the sampling of delays. At the delay model training step, different methods can be tested. In this work, a new approach using Machine Learning is presented, enabling to model more precisely delay distributions by automatically estimating the influence of different context-related factors based on what happened in the past. Indeed, many researches have shown dependencies between the observed delay and the context, e.g train type, hour, line, infrastructure, timetabling, capacity consumption, etc (Olsson and Haugland 2004; Abril et al. 2008). Other state-of-the art delay models are tested as benchmark.

Calibration of the delay distribution is usually done a posteriori by comparing simulated and observed values, however a priori calibration has important benefits. At first, it allows to identify precisely and easily bias in the probability distribution, while with the classical methodology it is difficult to separates errors in the distribution and errors in operating rules when results do not match. Moreover, in the case of simulation for robustness evaluation at station, observations are not concordant with the hypothesis of the simulation framework: some trains experience extreme delays, others are cancelled or modified during operations (e.g new schedule) or capacity may be constrained (e.g limited infrastructure). Observations and simulations can not be compared directly as they do not always include the same events. For instance, if a train is cancelled during operations, it will free infrastructure and reduce the opportunity of conflicts for surrounding trains. However, the robustness must be evaluated by taking into account every scheduled train, and in the simulation framework, every train might experience a delay. If reality and simulation outputs are compared, the scheduled train might have an initial delay or be impacted by other trains, but there are no observations to relate with.

Figure 2: data-driven simulation methodology



The main contributions are the use of Machine Learning to simulate delays and a new a priori calibration assessment methodology that evaluates the quality of the delay distribution before delays are sampled for the simulation. Different delay models with variable quality are tested, and the comparison of these models aims to highlight the impact of the quality of the delay modeling part on performance results.

4 Delay probabilities estimation

4.1 Classical delay models

This paper presents four alternatives to model delay distribution.

The first alternative consists in simulating delay scenarios with a negative exponential distribution. This method is very simple and doesn't require data as the distribution parameter just need to be set at the inverse of the mean delay value, or any other approximation of it. The mean value of the dataset containing all passenger train delays in minutes, excluding outliers (negative values are set to 0 and delays greater than 20 minutes are deleted) is used. This method is not realistic at all as all type of events are expected to follow the same

distribution, but it represents correctly the general behavior of train delays (high probability of small values, skewness, etc).

The second approach is similar but different train profiles are studied. Initial delays are also generated with a negative exponential distribution but the distribution parameter is depending on the type of train (high-speed, suburban,..) and the type of event (arrival and departure). This method is relevant when there are no available data but known statistics on the mean delay value. In this case, the parameters are set to the inverse of the mean value of the corresponding dataset. Table 1 gives main characteristics of the different delay types; it must be noticed that the average delay varies a lot.

The third approach computes empirical distribution based on historical records. The set is divided according to train types and event. For each of these data sets, a discrete probability function is built with the relative frequency of every delay value. This approach requires a database with observed delays, and some features (train types, event), but the calculations are easy. It is more realistic than the other method because it is built on historical observations and separate different cases. However, it doesn't consider more precise separations (line, stopping pattern, density of the traffic, type of day, peak hours, etc). It is possible to increase the number of clusters in order to consider more parameters, but it might affect the precision of the estimated empirical distribution as there will be less elements in each cluster.

The last one uses generalized linear models and is described in the next section.

4.2 A statistical learning approach

The methodology for delay modeling with Machine Learning is explained more precisely by Faverges et al. 2018a. It is based on three main steps: datasets creation, model training and goodness-of-fit assessment.

This approach relies on statistical properties of delay data (choice of a modeling distribution) and on learning aspects. It aims to estimate individual delay probabilities at station by taking into account the potential impact of other features. Moreover, calibration of these probabilistic predictions is evaluated based on the predictions.

Data collection

Historical records of train delays associated with a location and scheduled event time are collected for trains arriving at and departing from Montparnasse station. A data base is created for every train type (high speed, suburban and regional) and event type (arrival and departure). Relevant indicators are added and encoded to obtain a numerical set (e.g. origin, date, stopping pattern, type of day, arrival time, trip duration, etc).

Outliers are excluded from data sets. In practice, delays above a threshold are deleted. There are several reasons for this. At first large delays are rare and unpredictable. They do not have the same causes as small delays and add noise in data. Secondary, this paper focus on robustness to small delays, and simulating large delays will not reflect reality as in real-life large disturbances require specific actions to minimize their impact. Third, the Machine Learning approach used here is based on a maximum likelihood estimation, there is no need to optimize parameter based on unlikely and irrelevant observations. The truncation threshold depends on the type of event and type of trains: for arrivals, suburban trains are usually cancelled when they experience delays above 10 minutes while high-speed trains and intercity are maintained. The Montparnasse station has a high rate of punctual

Table 1: Data sets description

Set	size	truncation threshold	Mean value	Main features
High- speed arrivals	25900	20	3.08	stopping pattern, scheduled stopping time, type of day, time slot, traffic density (on line, at origin and destination), rolling stock
High-speed departures	28700	10	0.48	type of day, time slot, destination, traffic density in station, rolling stock
Suburban arrivals	38900	10	1.03	stopping pattern, scheduled stopping time, type of day, hour, traffic density (on line, at origin and destination), rolling stock
Suburban departures	40600	5	0.18	type of day, hour, destination, traffic density in station, rolling stock, duration
Regional and Intercity arrivals	11400	15	2.26	Origin, type of day, time, traffic density, rolling stock
Regional and Intercity departures	11500	7	0.45	type of day, time, destination, traffic density in station, rolling stock

departure trains due to its terminal station status, so a low threshold is enough. Beside extreme delays, some trains arrive in advance, in particular the high-speed trains. In this model, observations with negative values are set to zero. This is a strong assumption, but at this point, negative values are more complex to model and predict, and they are less relevant than positive delays for the robustness evaluation. Indeed, if a train arrive in advance and create a conflict with another train at the station, it is expected that the early train can wait until its schedule time, without creating new delay. These negative delays are rare (they concern usually only high-speed arrivals) and with small value (one or two minutes).

The data sets are described in table 1. The mean value is estimated among the truncated non negative values recorded in minutes. These data are collected over a year (summer 2017 to summer 2018), and they exclude days of major system failures and following days of recover (13 days), major scheduled works (10 days) and strikes days (32 days). Features are similar in the different sets, but they are processed differently. For instance, time slot is in hour for suburban trains as they have a high frequency, but it is a few hours for high-speed trains, the stopping pattern and scheduled stopping time make sense only for arrivals and not departures, etc.

Finally, each of these sets is separated into two parts: a training set that is used to build a model and a validation set to assess its goodnessof-fit.

Model training

A generalized linear model (GLM) is trained on each of the training sets (high speed arrivals, high speed departures, regional arrivals and regional departures). GLM are convenient in this case as they model a variable with a probabilistic distribution Faverges et al. 2018b; Faverges et al. 2018a. The prediction for each train is not a single value but the probabilities corresponding to every possible outcome. Different train types are separated to improve models performances by reducing heterogeneity: travel time instability has not the same causes for these different cases, and the same features may have different impact.

The R package GAMLSS is used to implement these models. It extends classic GLM by allowing a large variety of probabilistic distributions, by modeling multiple parameters simultaneously and enabling to truncate distributions (Stasinopoulos, Rigby, et al. 2007). On delay data from the Montparnasse station in Paris, the truncated negative binomial distribution (NBI) is chosen (Faverges et al. 2018a). It is the best compromise between complexity (number of distribution parameters to fit) and likelihood of the model. The model is displayed bellow with $\boldsymbol{Y}$ the delays, $\boldsymbol{X}$ the covariate matrix and $\boldsymbol{\beta_\mu}$ and $\boldsymbol{\beta_\sigma}$ estimated parameters. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the NBI distribution parameters.

$$\begin{cases} \boldsymbol{Y} & \sim \text{NBI}_{tr}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\ \ln(\boldsymbol{\mu}) & = \boldsymbol{X}\boldsymbol{\beta_\mu} \\ \ln(\boldsymbol{\sigma}) & = \boldsymbol{X}\boldsymbol{\beta_\sigma} \end{cases} \tag{1}$$

The figure 3 shows how the negative binomial distribution fits da ta. Observations are separated by type of train and events, and represented by the histograms. Parameters of the NBI are univariate maximum likelihood estimates of the true parameters, and corresponding probabilities are displayed with dots.

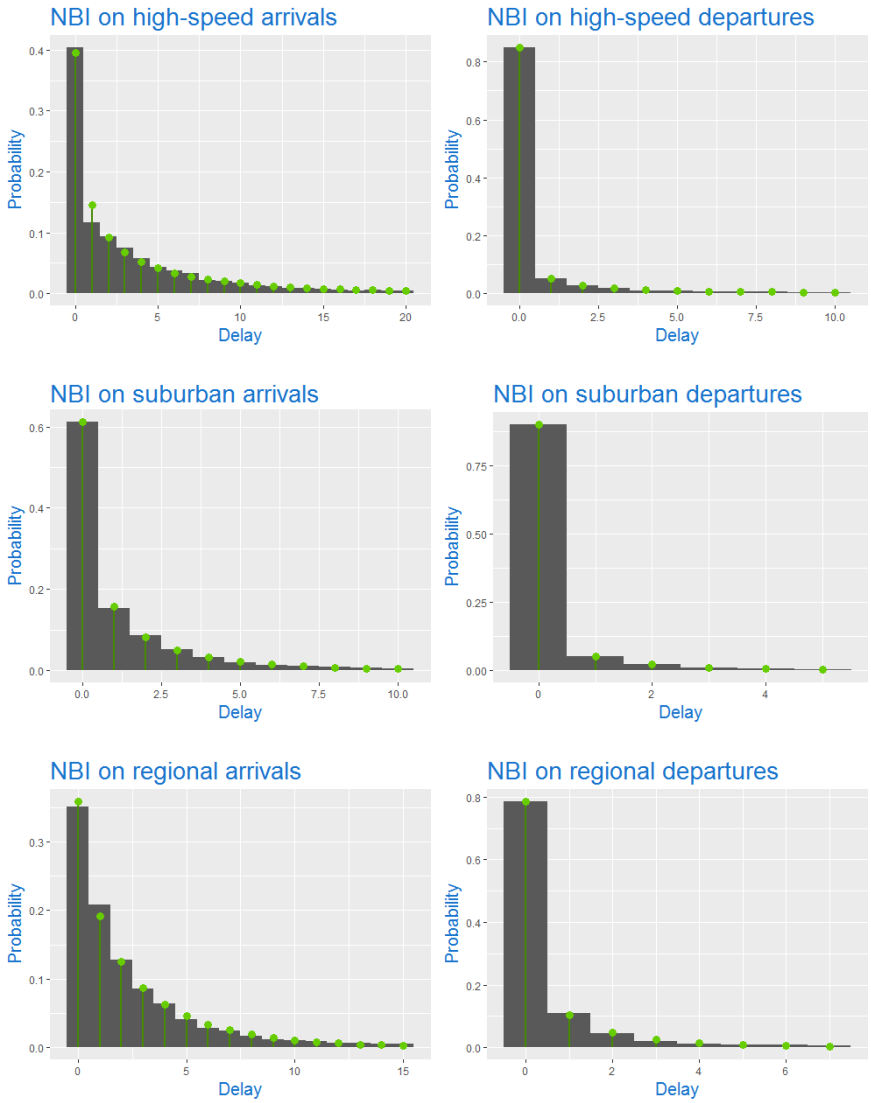
Model evaluation

As estimated individual probability mass functions are used to simulate delays, it is important to evaluate their quality and realism. However, usual residuals-based methods, like the mean absolute error, are not an option in this case because predictions and observations are not homogeneous (probability distribution and an integer).

This paper proposes to evaluate the quality of the model based on its calibration, ie the adequacy between estimated probabilities and observed rates of delay, at different points of the cumulative distribution functions. A graphical diagnostic of the calibration is done using grouping strategy: for a threshold given t, trains are sorted according to their estimated probability of having a delay higher than t and separated in g groups of similar predicted delay risk, then a calibration plot is obtained by displaying for each group the average estimated probability with the observed rate of delays in the group. A model can be considered calibrated when points are close to the diagonal.

This paper compares the different delay models only with calibration plots. An example of application of a statistical test to evaluate the significance of the deviations from the diagonal is given in Faverges et al. 2018a with the Hosmer-Lemeshow test. The plots are convenient as they are easily interpretable and allow to identify bias in the predictions (overestimation of risk for instance).

Figure 3: Negative binomial



5 Delay Propagation algorithm

This section presents the deterministic function that is applied on stochastic samples drawn from distributions computed previously. It aims to approximate the final delays based on the input scenario. These input delays (called *initial delays* here by opposition with *propagated* or *secondary* delays that are created in the station) are set to occur before the entrance of the station for arriving trains or at the platform for departing trains. This propagation function is an approximation of reality as during operations, many modifications are done on the original schedule (changes in the platform assignment, rolling stock management, human resources schedules, etc) based on human decisions.

For this preliminary study, a simple propagation algorithm is considered. It makes the strong assumptions that allocated paths are fixed and that the train sequence can be changed only if it is possible to maintain a train at its original schedule instead of delaying it. The goal is to study the reaction to delays without dispatching measures.

5.1 Station layout and constraints model

The infrastructure and constraints models are the ones implemented in the tool OpenGOV created by SNCF (cf section 3.1). There are two types of routes: arrival and departure, represented by an ordered succession of tracks. An arrival route is composed by an incoming track representing the entrance in the station and the beginning of a main line, three intermediate track sections and one platform track. A departure route has a platform track, three intermediate tracks and one outgoing track. Conflicting paths are defined as two paths that cannot be affected in a too short lapse of time, for instance if they share one or more tracks, or if they are crossing.

A solution of the platforming problem consists in the assignment of an arrival path and a departure path to each train. This assignment must respect rules, as described in section 3.1. The delay propagation algorithm has to take these constraints into account.

Minimal headways to respect between trains are set in OpenGOV according to the different cases of conflict and the station layout modeling: the type of train (high-speed, suburban, technical, etc), the type of event (arrival, departure, platform reoccupation, etc), the position of paths crossing (involved tracks), etc. The value associated with the different configurations corresponds to a security norm used for schedule conception. If a train is delayed, other trains on conflicting paths must wait the time necessary to ensure that the constraint is respected.

5.2 Algorithm

The algorithm is presented below. The notations are:

- $T = (t_1, ..., t_n)$ the list of trains sorted by schedule time $(h_{t_1}, ..., h_{t_n})$, and their corresponding simulated initial delays $(d_{prim,t_1}, ..., d_{prim,t_n})$
- The current delays $(d_{curr,t_1}, ..., d_{curr,t_n})$ correspond to the total delay of each train (initial and secondary). They are initialized at zero and then updated according to the delay propagation of other preceding trains and the initial delay of the corresponding train.

- For each ordered pair of trains (t, t') using conflicting paths, $cst_{t,t'}$ is the minimal headway to respect between the two trains. It depends on the type of train, the paths and the type of conflict
- For each $t \in T$, $CT_{prev(t)}$ is the list of previous conflicting trains with t, ie the list of trains t' that may impact t if they experience a delay higher than the scheduled buffer time $buffer_{t',t}$
- For each $t \in T$, $CT_{foll(t)}$ is the list of following conflicting trains, ie the list of trains t' that are impacted by t if t has a delay higher than the scheduled buffer time $buffer_{t,t'}$

The simulated initial delays are bounded by the truncation threshold, so they also produce bounded secondary delays. Moreover, only delays less than the maximal truncation threshold are considered to build $CT_{prev(t)}$ and $CT_{fol(t)}$.

In this simple algorithm, changes in the sequence are considered only if the train can be maintained at its original slot in order to cancel its secondary delay. These changes are possible only if they are compatible with all the trains originally scheduled before. For instance, two trains arriving at the station from the same track cannot be reordered, and it is naturally forbidden to exchange arrival and departure of the same train if it is delayed for the arrival and not the departure, but this is usually not a problem for a terminal station.

Algorithm 1 Propagation algorithm

Data: list of train $T = (t_1, ..., t_n)$ sorted by schedule time with their scheduled path, and their corresponding initial delays $(d_{t_1}, ..., d_{t_n})$

Result: Values of all secondary delays

initialization: Current delays are set to 0 $d_{curr,t} \leftarrow 0 \; \forall \; t \in T$

for $t \in T$ *sorted by schedule time* h_t **do**
  if $d_{curr,t} > d_{prim,t}$ **then**
    t has a secondary delay higher than its initial delay
    Test to verify if it is possible to maintain t at its original schedule time by changing the train sequence. It must be compatible with all the previous trains
    change = TRUE
    for $t' \in CT_{prev(t)}$ **do**
      if $h_{t'} + d_{curr,t'} < h_t + d_{prim,t}$ **then**
        if $h_{t'} + d_{curr,t'} + cst_{t',t} > h_t + d_{prim,t}$ **then**
          t' passes before t and the headway constraints is not fulfilled
          change = FALSE
        end
      else
        with its delay, t' passes after t. A change is the sequence may be possible
        if *t and t′ correspond to the arrival and departure of same train* **then**
          change = FALSE
        end
        if *t and t′ use the same platform or the same incoming track* **then**
          change = FALSE
        end
        if $h_t + d_{prim,t} + cst_{t,t'} > h'_t + d_{curr,t'}$ **then**
          the headway constraints is not fulfilled if t passes before t′
          change = FALSE
        end
      end
    end
    if *change = TRUE* **then**
      It is possible to change order of trains and maintain t at its original schedule with a potential initial delay but without secondary delays
      $d_{curr,t} \leftarrow d_{prim,t}$
    end
  else
    Current delay is set to initial delay
    $d_{curr,t} \leftarrow d_{prim,t}$
  end
  At this step, current delay of t is known. It is propagated to following trains
  for $t' \in CT_{foll(t)}$ **do**
    Secondary delay of t' is updated based on current delay of t
    $d_{curr,t'} \leftarrow \max\left(d_{curr,t'}, d_{curr,t} - \text{buffer}_{t,t'}\right)$
  end
end

6 Experiments

As described above, this paper presents four delay modeling alternatives for perturbations simulations. The differences between the results obtained by these approaches are studied by experimenting this methodology on a set of platforming solutions of the Montparnasse station. These solutions are the final schedules built by SNCF Reseau before operations. Four weeks are studied (the third week of the month of January, February, July and August). For the simulation part, 5000 iterations are done (delay simulation and propagation).

6.1 Differences between delay models

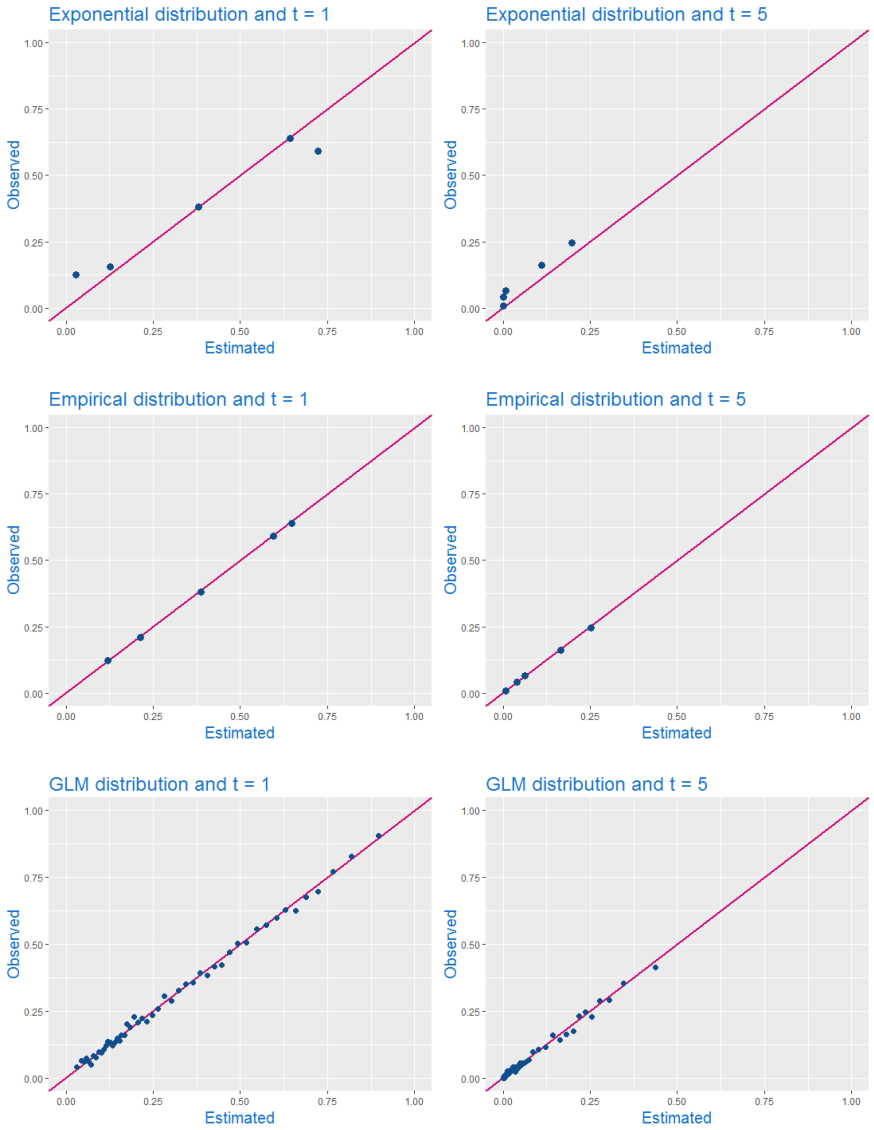
Three delays models are compared using calibration plots in Figure 4: the exponential modeling with one distribution per train and event type, the empirical distribution computed with historical records and the probabilities estimated with a GLM. The plots are build as described in subsection 4.2. A model is calibrated if points are close to the 45 degree: this means that estimated probabilities are concordant with observed delay frequency.

Two plots are displayed for each model: the first one to evaluate calibration of positive initial delay probability $\boldsymbol{P}(Y > 0)$ and the second one to evaluate the calibration of probabilities of delays greater than 5 minutes $\boldsymbol{P}(Y \geq 5)$.

For the exponential and the empirical models, there are only 6 groups possible as estimated probabilities are the same among different clusters *train type/event type*. The GLM model estimates delay probabilities using more features, so the range of predicted probabilities is larger and predictions are individual. It is visible on the graph since points modeling groups spread on the diagonal (50 groups are used). The model is more discriminant because it successfully recognize more punctual trains with low estimated probability from more delayed ones with higher estimated probability. It is also well calibrated.

The empirical model is very well calibrated as points are really close to the identity line. However, these probabilities are not precise, they only have a few values possible. The exponential model shows deviations between observations and estimations. In particular, it overestimates the large values of $\boldsymbol{P}(Y > 0)$ (points under the line) and underestimates small values (points over the line). $\boldsymbol{P}(Y \geq 5)$ is slightly underestimated for all clusters. Samples drawn with this model might differ from reality as certain trains are systematically more (or less) delayed that what is observed i dta.

Figure 4: Calibration plots



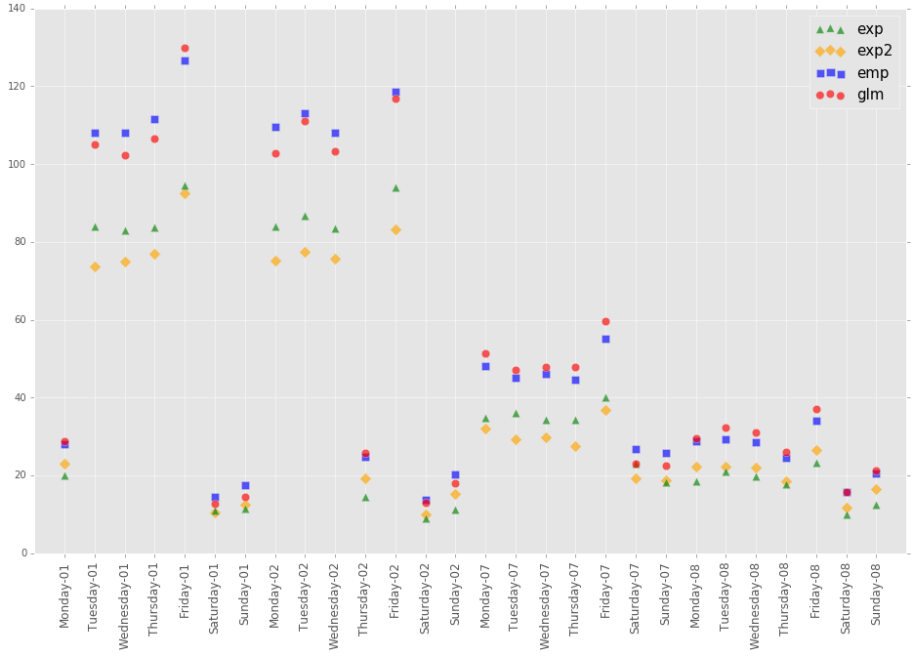
6.2 Propagation results

For each instance and each input delay scenario, two indicators are studied to compare the different initial delays modeling: the number of trains with a positive secondary delay created at station and the mean value of these positive secondary delays. They are computed after the propagation of the initial delays and averaged over the N iterations, considering

only passenger trains on four weeks of January, February, July and August. *exp* stands for the simplest exponential model with an unique delay distribution for all trains, *exp2* for the model with one distribution per cluster, *emp* for the empirical distributions per cluster based on time-stamps data and *GLM* for the learning model.

Results are given on the following plots.

Figure 5: number of secondary delayed passenger trains

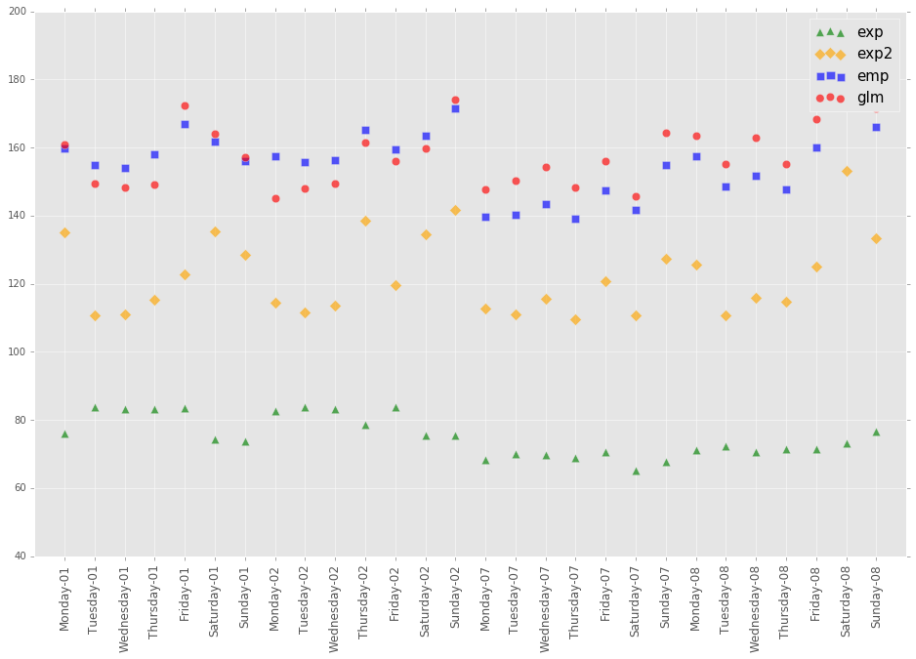


Considering the number of trains with positive secondary delay on Figure 5, all delay models provide highly correlated values with a redundant pattern of less delayed trains with the exponential models (especially *exp2*) and a larger number of secondary delayed trains with the *GLM* model.

Regarding the average positive secondary delays on Figure 6, strong deviations are observed between models. The *exp* model underestimates strongly the mean delay. This is not surprising, as arrivals delays are larger than departure delays in reality but modelled with the same distribution here, this model tends to simulate more initial delays with smaller values than the other models. They propagate less longer. Despite a correct calibration, the *exp2* model also shows important deviations with the other models, probably because it underestimates the probability of larger delays (cf Figure 4).

Finally, the average delay is close between the models *emp* and *glm* on all instances with differences of only a few seconds. They are both calibrated, and differences in discrimination doesn't have a visible impact on the average positive secondary delay.

Figure 6: Average secondary passenger train delay in seconds



7 Discussion

This work presents preliminary results on the perspective of delay modeling with Machine Learning to evaluate and improve robustness of operations at station. In particular, it focuses on the impact of calibration of delay modeling and expresses the difficulty to calibrate operating rules.

A priori calibration: it consists in assessing the goodness of fit of the perturbations distributions before the propagation algorithm. This approaches has several benefits:

- A posteriori calibration requires to compare results with actual observations that includes several outliers (large delays, but also cancelled trains that are not observed and may impact results). These outliers are not relevant for a robustness study, which focuses on small deviations of input parameters.
- A robust solutions must absorb delays with a limited use of dispatching. However, in reality, multiple changes are made on the schedule and the propagation is performed differently. In particular for the routing phase, alternative paths are preferred to propagation during operations. Calibrating probabilities based on results that are obtained with different processes may affect the results.
- Finally, using a priori calibration is promising to develop and test new delay propagation and dispatching strategies.

Delay propagation This paper also raises the issue of realism of propagation algorithm. This algorithm aims to represent real operations, but in this case of delay propagation in station, there are several limits to its realism:

- During operations, it is common to change path assignment, and even platform to avoid delay propagation. However, this is unrealistic to use such algorithm in a simulation framework as this is a complex decision problem. Moreover, it also must be avoided for a robustness study as a solution is not robust if infrastructure managers must perform multiple changes to the schedule in order to avoid delay propagation.
- The security constraints used in this study are sometimes too conservative. They correspond to conception constraints and must be respected when the schedule is conceived. However, in certain cases during operations, the required time between the two trains is less than the security constraint for conception, and the second train can pass before the constraint is respected.

Perspectives

- This methodology should be tested on platforming solutions with different level of robustness to evaluate more precisely the impact of input calibration. This paper shows that a bad calibration can lead to false magnitude in results. Working on same solutions of the same day would help to see the relative impact of the delay distribution when they are compared based on their robustness.
- In addition of delay distribution, more work should be done on propagation algorithm and operating rules calibration.
- Differences between delay distribution modeling should be studied at a more microscopic level, for instance to detect the differences due to systematic delays (trains that systematically experience secondary delays at station, useless buffer times,...). This will help to detect robustness defaults in solution.
- The delay modeling part could be improved with more precise data: delays are recorded in minute in this data set, this lack of precision add noise in results. Moreover, other modeling strategies than GLM should be tested for individual probabilities, like for instance Random Forests.

8 Conclusions

This paper presents a simulation methodology for robustness evaluation at station using statistical techniques (Generalized Linear models to estimate delay distributions according to the context and calibration plots to assess the goodness-of-fit) to provide a more realistic delay model that doesn't require a posteriori calibration. A robustness evaluation framework is used, characterized by truncated delay distributions and simple dispatching measures. A priori calibration is suitable in this case, as these assumptions do not totally reflect reality (large delays require specific dispatching that is not modeled here and trains are sometimes cancelled), comparing simulation results with observations to assess calibration might lead to bias.

The generalized linear model is compared with three other delay modeling techniques (two exponential models and one empirical distribution). The calibration assessment shows that the GLM and the empirical distribution are both well calibrated, unlike the exponential model that shows slight deviations. The GLM is also more precise and achieve to discriminate better the most punctual trains from the most delayed one while other models use the same probabilities among different clusters.

These modeling approaches are used for simulation of operations at station on 28 platforming problem solutions. Based on the number of trains experiencing secondary delays and the value of the average positive secondary delay, the empirical distributions and the GLM distributions give similar results while other models show strong deviations.

Acknowledgements

This work is conducted as part of a CIFRE PhD convention for an industrial agreement between SNCF Reseau and the CEDRIC laboratory - CNAM.

Authors thank Hajar Taleb for her precious help on OpenGOV tool and Rémi Parel for his knowledge on operations at Montparnasse station. Authors also wish to acknowledge the help provided by Antoine Robin, who worked on a preliminary version of the propagation algorithm during his internship at SNCF Réseau.

References

Abril, M et al. (2008). "An assessment of railway capacity". In: *Transportation Research Part E: Logistics and Transportation Review* 44.5, pp. 774–806.

Armstrong, John and John Preston (2017). "Capacity utilisation and performance at railway stations". In: *Journal of Rail Transport Planning & Management* 7.3, pp. 187–205.

Bergström, Anna and Niclas Krüger (2012). "Modeling Passenger Train Delay Distributions: Evidence and Implications for Valuation". In: *The 5th International Symposium on Transportation Network Reliability (INSTR2012), Hong Kong, China, December 18-19, 2012*. Pp. 61–61.

Briggs, Keith and Christian Beck (2007). "Modelling train delays with q-exponential functions". In: *Physica A: Statistical Mechanics and its Applications* 378.2, pp. 498–504.

Büker, Thorsten and Bernhard Seybold (2012). "Stochastic modelling of delay propagation in large networks". In: *Journal of Rail Transport Planning & Management* 2.1-2, pp. 34–50.

Caprara, Alberto et al. (2010). "Robust train routing and online re-scheduling". In: *OASIcs-OpenAccess Series in Informatics*. Vol. 14. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Carey, Malachy (1999). "Ex ante heuristic measures of schedule reliability". In: *Transportation Research Part B: Methodological* 33.7, pp. 473–494.

Carey, Malachy and Sinead Carville (2000). "Testing schedule performance and reliability for train stations". In: *Journal of the Operational Research Society* 51.6, pp. 666–682.

Cui, Yong, Ullrich Martin, and Weiting Zhao (2016). "Calibration of disturbance parameters in railway operational simulation based on reinforcement learning". In: *Journal of Rail Transport Planning & Management* 6.1, pp. 1–12.

Faverges, Marie Milliet de et al. (2018a). “Estimating Long-Term Delay Risk with Generalized Linear Models”. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 2911–2916.

— (2018b). “Modelling passenger train arrival delays with Generalized Linear Models and its perspective for scheduling at main stations”. In: *ICRE*. IET.

Goverde, Rob MP (2005). “Punctuality of railway operations and timetable stability analysis”. In:

Harrod, Steven, Georgios Pournaras, and Bo Friis Nielsen (2018). “Distribution Fitting for Very Large Railway Delay Data Sets with Discrete Values”. In: *Trafikdage 2018*.

Koutsopoulos, Haris and Zhigao Wang (2007). “Simulation of urban rail operations: application framework”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2006, pp. 84–91.

Kroon, Leo G, H Edwin Romeijn, and Peter J Zwaneveld (1997). “Routing trains through railway stations: complexity issues”. In: *European Journal of Operational Research* 98.3, pp. 485–498.

Landex, Alex and Otto Anker Nielsen (2006). “Simulation of disturbances and modelling of expected train passenger delays”. In: *Timetable Planning & Information Quality. Delft University of Technology, WIT Press Publishing, The Netherlands*, pp. 85–93.

Larsen, Rune et al. (2014). “Susceptibility of optimal train schedules to stochastic disturbances of process times”. In: *Flexible Services and Manufacturing Journal* 26.4, pp. 466–489.

Olsson, Nils OE and Hans Haugland (2004). “Influencing factors on train punctuality—results from some Norwegian studies”. In: *Transport policy* 11.4, pp. 387–397.

Sels, Peter et al. (2014). “The train platforming problem: The infrastructure management company perspective”. In: *Transportation Research Part B: Methodological* 61, pp. 55–72.

Stasinopoulos, D Mikis, Robert A Rigby, et al. (2007). “Generalized additive models for location scale and shape (GAMLSS) in R”. In: *Journal of Statistical Software* 23.7, pp. 1–46.

Wen, Chao et al. (2017). “Statistical investigation on train primary delay based on real records: evidence from Wuhan–Guangzhou HSR”. In: *International Journal of Rail Transportation* 5.3, pp. 170–189.

Yuan, Jianxin (2006). “Stochastic modelling of train delays and delay propagation in stations”. In: