Estimating long-term delay risk with Generalized Linear Models

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

Estimating long-term delay risk with Generalized Linear Models

1st Marie Milliet de Faverges
CEDRIC - CNAM
SNCF Reseau - DGEX Solutions
Paris, France
marie.milliet-de-faverges@reseau.sncf.fr

2nd Giorgio Russolillo
CEDRIC - CNAM
Paris, France
giorgio.russolillo@cnam.fr

3rd Christophe Picouleau
CEDRIC - CNAM
Paris, France
christophe.picouleau@cnam.fr

4th Boubekeur Merabet
SNCF Reseau - DGEX Solutions
Saint Denis, France
boubekeur.merabet@reseau.sncf.fr

5th Bertrand Houzel
SNCF Reseau - DGEX Solutions
Saint Denis, France
bertrand.houzel@reseau.sncf.fr

***Abstract*—This paper presents an original methodology to estimate delay risk a few days before operations with generalized linear models. These models represent a given variable with any distribution from the exponential family, allowing to compute for any subject its own probability distribution according to its features.**

This methodology is applied on small delays (less than 20 minutes) of high-speed trains arriving at a major french station. Several distributions are tested to fit delay data and three scenarios are evaluated: a single GLM with a negative binomial distribution and two two-part models using both a logistic regression as first part to compute the probability of arriving on time, and a second part using a negative binomial or a lognormal distribution to obtain the probabilities associated with positive delay values.

This paper also proposes a validation methodology to assess the quality of these probabilistic predictions based on two aspects: calibration and discrimination.

***Index Terms*—Railway punctuality, Generalized Linear Models, Delay modeling, zero-inflated data**

I. Introduction

Demand in public transportation has strongly increased these past years and important investments have been made to support this growth. However, for the railway sector new resources are expensive and long to deploy. A strong utilization of available resources presents benefits as more trains can be scheduled, but the service is less reliable and the average delay is higher. This is the trade-off between capacity and reliability [1].

An efficient management of capacity is necessary to simultaneously provide a quality service and respond to railway demand. However, network saturation and potential delays make scheduling processes very challenging. Delays impact planning feasibility and real-time rescheduling is inevitable. This phenomenon is particularly pronounced in stations where most of delays occur and propagate, due to route conflicts, passenger activity or synchronization of connections (crews, rolling stock or passenger). Strategies exist to limit this propagation, for instance by imposing buffer times between two successive uses of the same resource. However, margins are limited by available capacity and buffer times need to be affected wisely.

As a response to congestion issues and with recent technological advances, large investments have been made to provide tools to support decisions for traffic management [2]. In particular, different traffic monitoring devices and automatic data collection tools have been developed. Machine learning methods can be used for analyzing these data to explain a part of traffic variability. Understanding this variability can help to increase level of service, for instance with real-time rescheduling tools, passenger information systems or simulation methods.

This paper presents an original use of Generalized Linear Models on historical railway data to estimate delay risk.

This paper is structured as follow. Section 2 exhibits a rapid overview of literature on delay prediction. Section 3 describes the problem and the case study. Section 4 presents theory about methods that are used here. Experiments are depicted in section 5 and results are given in section 6. Section 7 addresses considerations about obtained results, limits and perspectives. Conclusions are provided in the final section.

II. Related Work

There has been an important number of researches dedicated to travel time and delay analysis in public transportation, and in particular for their prediction. Two different type of study using Machine Learning and data mining techniques have been identify based on prediction horizon: short-term prediction and long-term prediction [3].

The first one corresponds to an analysis at operational level. It is fed with real-time data, and it aims to estimate delay at next stops knowing the current delays through the network. Predictions can be used for rescheduling or passenger information. Peters et al. [4] propose a passenger train delay prediction with neural networks aiming to limit

delay propagation with intelligent real-time timetable monitoring. Pongnumkul et al. [5] estimate arrival time more accurately with moving average and k-nearest neighbors algorithms than with a simple translation of the current delay. Kecman and Goverde [6] apply linear regression, decision trees and random forests in order to predict running and dwelling times considering current train position and traffic information. Oneto et al. [7] present extreme learning machines used on historical and exogenous data like weather records to predict the delay that will affect a train at next checkpoints with respect to its delay at last visited points and to the delays of other trains running over the same section.

On the other hand, there are long-term delay predictions which are conducted at a tactical level (few months to days before operations) or even at a strategic level. These studies usually utilize timestamps of trains at specific points of the network and aim to discover pattern in delays. They have interesting perspectives for service improvement, e.g. more robust planning, new dispatching strategies or investments support. Markovic et al. [3] propose a prediction model of arrival train delays with support vector regression in order to have a better understanding of the relationship between infrastructure and delays. Cerreto et al. [8] apply k-means clustering to identify different delay profiles, providing new insights for managerial decisions.

III. Case study

A. Proposed Framework

This work aims to support tactical decisions for main stations management, and in particular to improve robustness of solutions of the platforming problem by integrating actual delay risk at least one day before operations. This problem consists in routing trains through station and affecting them platforms. An inadequate solution can have a strong impact on punctuality [9].

From an operations research perspective, a robust solution can be defined as a solution which remains feasible in real conditions when small disturbances occur. For the platforming problem, a good schedule must produce limited delay propagation during operations [10]. For this reason, this study analyses only delays below 20 minutes: they have a significant impact on schedule feasibility, they occur frequently and they produce consecutive delays [10]. Larger delays are less predictable and have different causes: they add noise in the data set. Moreover, they are not relevant for the routing problem: dispatching actions will be necessary to route the delayed train [11].

Early trains are considered punctual in this analysis (negative delays are set to zero). This is a strong assumption as they might have consequences on the scheduling. However, this study focuses on delayed trains in a fist place, and restricting the variable domain to non-negative integers allows to use state-of-the art distributions.

B. Context

This study takes place at Montparnasse station in Paris, France. This station has a complex infrastructure and hosts about 800 arriving and departing trains per day, making the routing task difficult. Different services (high-speed, suburban, regional trains) are studied with separate models of delay risk estimation. This avoids to merge heterogeneous data in the same analysis. This paper presents a model for high-speed trains arrival delays.

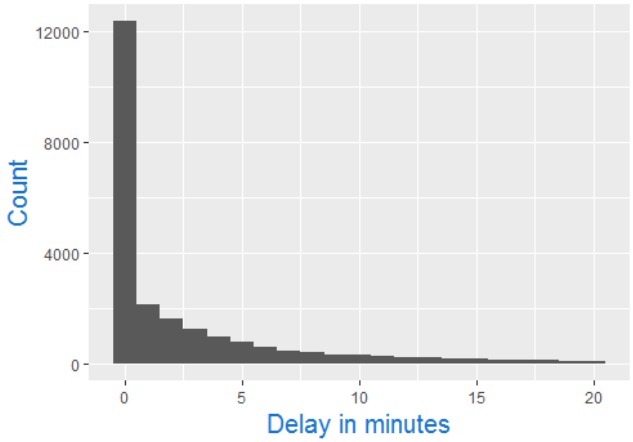

Fig. 1: High-speed arrival train delays

C. Data

Observations and most of explanatory variables are extracted from historical track occupation data. These data contain information for each passing train such as the train number, the observed delay in minutes or the observation time and the date. They were collected at Montparnasse station in Paris and at every stations of the high-speed network to reconstitute each train's journey, e.g. its origin, its stopping pattern or density of the traffic during its trip. Other data sources are used, such as weather data or school holidays calendar.

Data of all high-speed trains from the 1st July 2016 to the 30th June 2017 were extracted, representing approximately 27 500 trains. Observations with an unusual stopping pattern (defined as less than 0.3% of the occurrences) were deleted because they might correspond to errors in the data or exceptional circumstances. Trains with delays above 20 minutes were also deleted. The remaining 23 000 trains were used for this study. Moreover, trains with negative delay were set to zero. Corresponding delay distribution is displayed in Fig. 1. It is heavily right skewed with a peak at zero.

IV. Method choice

A. Generalized Linear Models

Generalized linear models (GLMs) extend linear regression by allowing to model non-normal response with eventually non-constant variance. They have three components [12]:

- A random component, with the response variable $\boldsymbol{Y}$ and its distribution from the exponential family.
- Linear predictors : $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ with $\boldsymbol{X}$ the covariate matrix and $\boldsymbol{\beta}$ the parameter vector.

- A link function g that relates the random component to the predictors : $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ with $\boldsymbol{\mu}$ the distribution parameter vector.

A generalized linear model can be written as :

$$\begin{cases} \boldsymbol{Y} & \sim \text{Exp}(\boldsymbol{\mu}) \\ g(\boldsymbol{\mu}) & = \boldsymbol{X\beta} \end{cases} \quad (1)$$

These models estimate the distribution parameter μ of each observation. Regression coefficients vector $\boldsymbol{\beta}$ is determined by maximum likelihood. The response variable can have various shapes according to the distribution choice : GLMs can model binary data, count-data and continuous data on different domains, e.g. real or positive.

B. Two-part models

Delays are known to be well fitted by distributions from the exponential family, like the negative-exponential, Weibull or lognormal distributions [13], [14].

In this study, delays are non-negative integers with a peak at zero as many trains arrive on time. This clumping makes state-of-the-art distributions inappropriate to model train arrival delays with a GLM. Indeed, several density functions, including lognormal and Weibull, are not defined in zero. Even with discrete distributions, GLMs can have a lack of fit due to a disproportionate rate of zero values [15].

A classical method to handle these zero-inflated data is to use two-part models [12], [15]. The first part is a binary model, like a logistic regression, for the dichotomous event of having zero or positive observations. The second part fits values above zero with a continuous or discrete distribution. It is trained only on positive observations(without zero values). A global cumulative distribution function is obtained by combining predictions of the two parts.

Let $\boldsymbol{Y} = (y_1, .., y_n)$ be the observations vector and $\boldsymbol{X}$ the covariate matrix.

Then for $t \geq 0$:

$$\mathbb{P}[y_k \leq t|\boldsymbol{x_k}] = \pi_k + (1 - \pi_k)\mathbb{P}[y_k \leq t|y_k > 0, \boldsymbol{x_k}] \quad (2)$$

The first model estimates $\pi_k = \mathbb{P}[y_k = 0|\boldsymbol{x_k}]$ and $\mathbb{P}[y_k \leq t|y_k > 0, \boldsymbol{x_k}]$ is determined with the second model for each y_k and each t.

C. Validation

Goodness-of-fit assessment of probability prediction models is more difficult than punctual prediction models which are usually evaluated by means of the difference between the predicted values and the observed outcomes (residuals). For probability predictions, the model outcome and the response variable are not homogeneous: on one hand there is a probability and on the other hand a real non-negative observation.

An option is to evaluate the predicted cumulative distribution functions at several time-points. A binary framework is used : a success (observation equal to 1) corresponds to an arrival after t minutes and the prediction is the probability of having a delay greater than t minutes. In this framework, goodness-of-fit is usually assessed with discrimination and calibration analysis [16].

1) Calibration: it refers to model's ability to estimate probabilities that are consistent with observed rate of events. It can be tested either graphically or statistically for binary responses. In both cases, subjects are first sorted by their predicted probability of success and separated in g equal-sized groups, then predictions and observations are compared among each group.

A calibration plot is obtained by displaying for each group of subjects the average estimated probability with the actual proportion of successes. For a well-calibrated model, the plot should be close to the 45-degree line.

Model goodness-of-fit is usualy assessed by the Hosmer-Lemeshow test [17]. The Hosmer-Lemeshow statistic is given by :

$$C_g = \sum_{i=1}^{g} \frac{(O_{1,i} - E_{1,i})^2}{E_{1,i}} + \frac{(O_{0,i} - E_{0,i})^2}{E_{0,i}}, \quad (3)$$

where $O_{1,i}$ and $E_{1,i}$ (resp. $O_{0,i}$ and $E_{0,i}$) are the observed and predicted number of successes (resp. failures) in group i. Under the null hypothesis that the model fits the data, C_g follows a χ^2 distribution with $g-2$ degrees of freedom [18]. A well calibrated model is expected to have a non-significant p-value, e.g. greater than 0.05, which does not lead to reject the null hypothesis.

2) Discrimination: It refers to the ability to distinguish successes from failures based on the predicted probabilities. It is usually measured with the area under the ROC curve (AROC) [16]. Indeed, it evaluates the likelihood for a positive event of having a higher predicted probability of success than a negative event [17]. If $AROC = 0.5$, the model doesn't discriminate. Discrimination increases with $AROC$, and values greater than 0.7 are in general considered as satisfactory [17].

Discrimination and calibration are two separate measures. Indeed, a model that predicts the same probability to all subjects can be well calibrated but won't discriminate at all, while a model with systematic errors on estimated probabilities can discriminate subjects with different outcomes as predictions ranks are preserved.

As predictions are expected to be used to adapt planning in a major station, both aspects are required. Indeed, the different train types are studied with separate models. Probabilities must be calibrated to allow using them simultaneously to take decisions, otherwise a type of train might be privileged due to bias in estimations. Discrimination is important to develop robustness strategies: residual capacity can be used where it is most needed only if the model distinguishes accurately different levels of risk.

V. Experiments

The purpose of this experiment is to compare performances of single GLM and two-part GLM on small delays data.

All experiments are performed with the R package GAMLSS [19], which allows fitting data with a large variety of distributions, including truncated ones, and to estimate several parameters of a distribution simultaneously. The

package also has functions to fit distribution to data and to perform step-wise feature selection, which are used here.

The dataset is randomly divided in a train set with 17 200 observations and a test set with 5 800 observations. Models are trained on the train set and their performances are also evaluated on the test set in order to verify their ability to generalize to unknown data.

A. Distribution choice

A distribution has to be chosen to model the variable "delay" in a GLM framework. This distribution can be either discrete or continuous. In a single-part model, the distribution fits all values, otherwise it fits only the positive ones.

For each scenario, distributions from the package are tested. They all are truncated on the right at 20 minutes. Distributions are compared and sorted based on their Akaike criterion in Table I. Part 1. compares discrete distributions applied on the full train set. Part 2. also compares discrete distributions but only on positive values of the train set. Distributions are truncated at 0 for these experiments. Part 3. contains results on continuous distributions defined for positive values.

TABLE I: Distributions goodness-of-fit

Part 1. Discrete distributions on full data set		
Distributions	***Parameters***	***AIC***
Zero Inflated Poisson Inverse Gaussian	3	64 950
Zero Adjusted Negative Binomial	3	64 980
Negative Binomial	2	65 100
Delaporte	3	65 100
Sichel	3	65 100
Zero adjusted logarithmic	2	65 390
Waring	2	66 470
Yule	1	66 590
Poisson inverse gaussian	2	67 030
Geometric	1	72 800
Zero inflated Poisson	2	79 220
Poisson	1	123 260
Part 2. Discrete distribution for positive delays		
Distributions	***Parameters***	***AIC***
Sichel	3	41 140
Waring	2	41 150
Delaporte	3	41 160
Negative Binomial	2	41 180
Geometric	1	41 330
Yule	1	42 770
Poisson	1	54 410
Part 3. Continuous distribution for positive delays		
Distributions	***Parameters***	***AIC***
Box-Cox Power Exponential	4	39 850
Box-Cox-Cole-Green	3	40 900
Box-Cox t	4	40 900
Generalized Inverse Gaussian	3	40 930
Lognormal	2	41 470
Gamma	2	42 360
Weibull	2	42 526
Negative Exponential	1	42 680

B. Models

Many distributions have similar results based on Akaike criterion. Simpler distributions, such as Poisson or Negative Exponential have the worst results. Since in practice algorithms for fitting distribution with three parameters require a huge computational burden and often do not converge, only distributions with one or two parameters are considered here.

Three models are tested :

- Model 1: a single count model with a negative binomial distribution
- Model 2 : a two-part model with a logistic regression and a left-truncated negative-binomial distribution
- Model 3 : a two-part model with a logistic regression and a lognormal distribution

In these three cases, distributions are right-truncated as there are no values greater than 20. Second parts of two-part models are trained on positive values of the train set, representing about 8 000 subjects.

C. Feature selection

The dataset contains many redundant features and non-informative predictors. A stepwise procedure is used to obtain a relevant features subset with a greedy search method. The initial model is the null model with an empty feature subset. For m potential variables, the two following steps are iterated until the criterion stops decreasing:

- m models are generated by adding one new feature to the subset or by deleting one of the current subset.
- The model with the smaller criterion is kept and the subset is updated.

The criterion is a function of the log-likelihood penalized by the size of the subset : $C = pk - 2l$ with l the model log-likelihood, p the penalty and k the size of the subset.

In this work, the procedure is iteratively applied to all distribution parameters if there are several until convergence and k has been set to 3 according to [20].

Original dataset has 110 potential explanatory variables. After the feature selection procedure, Model 1 has 36 features for μ and 35 features for σ. Model 2 and Model 3 have a subset of 43 explanatory variables for their fist part with the logistic regression. Second part of Model 2 uses 18 features to model μ and 17 to model σ. Second part of Model 3 uses 34 features for μ and 10 features for σ.

VI. Results

Models are evaluated as follow : for each time-point t from 1 to 20 and for each observation, $\mathbb{P}[y \geq t|\boldsymbol{x}]$ is calculated, where y is the delay and $\boldsymbol{x}$ the covariate vector. In Model 1, this probability is directly estimated from the cumulative distribution function of the modeled probability distribution given the parameter estimates. In Model 2 and Model 3 it is estimated as described in equation (2).

$\mathbb{P}[y \geq t|\boldsymbol{x}]$ is then considered as the estimated probability of success of a binary GLM. Calibration and discrimination are tested and results are given in the following tables. For each time-point t and each proposed model, the p-value of the Hosmer-Lemeshow test and the area under the ROC curve are calculated. The number of groups g used for the Hosmer-Lemeshow test is computed according to recommendations given by Paul and Lemeshow [21] to standardize the power of the test when the sample size or the successes rate are

TABLE II: Model comparison on train set

parameters			Model 1 NBI		Model 2 LR - NBI		Model 3 LR - LOGNO	
t	*f*	*g*	*p*	*ROC*	*p*	*ROC*	*p*	*ROC*
1	46	1190	**0.91**	0.66	**0.63**	0.66	**0.63**	0.66
2	37	1190	**0.34**	0.66	**0.90**	0.65	0.00	0.65
3	30	1190	**0.24**	0.66	**0.16**	0.66	**0.05**	0.66
4	25	1064	**0.52**	0.67	**0.49**	0.66	**0.48**	0.66
5	20	875	**0.40**	0.67	**0.59**	0.67	**0.52**	0.67
6	17	720	**0.58**	0.68	**0.05**	0.68	**0.46**	0.68
7	14	608	**0.97**	0.69	**0.87**	0.68	**0.34**	0.69
8	12	520	**0.84**	0.69	**0.39**	0.68	**0.29**	0.69
9	10	442	**0.07**	0.69	**0.94**	0.69	**0.15**	0.69
10	9	379	**0.23**	0.69	**0.41**	0.69	**0.18**	0.69
11	7	325	**0.16**	0.70	0.02	0.69	0.01	0.69
12	6	274	0.02	0.70	**0.24**	0.69	0.01	0.69
13	5	230	**0.42**	0.70	0.37	0.70	0.00	0.70
14	4	190	**0.74**	0.69	**0.14**	0.69	0.00	0.69
15	4	152	**0.20**	0.69	0.00	0.68	0.00	0.69
16	3	117	**0.20**	0.68	0.00	0.67	0.00	0.66
17	2	89	**0.62**	0.67	0.03	0.64	0.00	0.65
18	2	63	**0.06**	0.64	0.01	0.61	0.00	0.60
19	1	37	**0.08**	0.57	0.04	0.55	0.00	0.54
20	<1	18	**0.18**	0.50	0.02	0.51	0.00	0.50

TABLE III: Model comparison on test set

parameters			Model 1 NBI		Model 2 LR - NBI		Model 3 LR - LOGNO	
t	*f*	*g*	*p*	*ROC*	*p*	*ROC*	*p*	*ROC*
1	46	133	**0.15**	0.65	**0.19**	0.65	**0.19**	0.65
2	37	133	0.00	0.66	**0.27**	0.65	0.00	0.65
3	30	133	**0.08**	0.66	**0.06**	0.66	0.00	0.66
4	24	133	**0.64**	0.66	**0.35**	0.66	0.00	0.66
5	20	133	**0.53**	0.66	**0.14**	0.66	0.04	0.66
6	17	133	**0.62**	0.67	0.01	0.67	**0.14**	0.67
7	15	133	**0.37**	0.67	0.01	0.67	**0.40**	0.68
8	12	133	0.04	0.68	**0.36**	0.68	**0.33**	0.68
9	11	133	**0.46**	0.68	**0.25**	0.68	**0.21**	0.69
10	9	130	**0.38**	0.68	0.04	0.69	**0.08**	0.69
11	8	109	0.02	0.69	**0.37**	0.69	**0.60**	0.69
12	6	92	**0.16**	0.70	**0.25**	0.69	**0.27**	0.70
13	5	76	**0.28**	0.69	**0.05**	0.70	0.00	0.68
14	4	61	**0.11**	0.70	**0.31**	0.70	**0.19**	0.70
15	4	50	**0.46**	0.71	0.00	0.70	**0.15**	0.70
16	3	40	**0.13**	0.69	**0.19**	0.70	**0.05**	0.69
17	2	29	**0.32**	0.69	0.03	0.67	**0.06**	0.66
18	1	20	**0.36**	0.67	**0.34**	0.65	**0.37**	0.64
19	1	12	**0.24**	0.61	**0.36**	0.55	**0.56**	0.56
20	<1	10	**0.44**	0.50	**0.44**	0.52	**0.55**	0.50

varying. f represents the percentage of trains arriving with a delay greater than t.

A model is considered calibrated with a non-significant p-value (greater than 0.05) and the greater is the area under the ROC curve, the better the model discriminates. However, correct models might still have small p-value. Diagnostic has to be done with a calibration plot to evaluate how strong are the deviations between observations and estimations.

For instance the null hypothesis is rejected for Model 1 and Model 3 for $t = 2$ on the test set. Fig. 2 shows corresponding calibration plots for $g = 40$. Model 3 is not well calibrated as the plot is under the diagonal : it overestimates the probability of having a delay greater than 2 minutes. On the opposite,

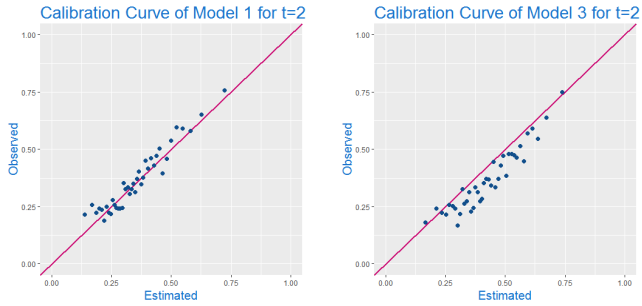

Fig. 2: Graphical diagnostic of calibration

Model 1 seems to be correctly calibrated.

VII. Discussions

A. Model choice

GLMs constitute a promising alternative to model train delays: they are based on probabilistic predictions instead of punctual values like state-of-the-art approaches. With this methodology, it is possible to evaluate the risk associated with any delay value.

Two options are presented here with classic GLM and two-part GLM. A single part model is easier to use with only one model to fit on the full data set but two-part models have some benefits too. At first they allow to represent train delays with state-of-the-art continuous distributions. These distributions, e.g. Weibull, Gamma or Lognormal, can't fit positive mass in zero, excluding modeling delays with a single GLM. Then a two-part model is simpler as it requires smaller feature subset for each parameter estimator, and the second part is trained on a subset of data. This is interesting as estimating a two-parameters GLM may take a while to converge, especially on large datasets. Therefore, feature selection procedure takes much more time for Model 1 than for Model 2.

B. Results analysis

When t increases, the proportion of trains with a delay greater than or equal to t, decreases strongly. Predicted probabilities become small and their range is narrow. This can cause borderline effects when calibration is estimated because the number of successes per group is limited and standardization of the Hosmer-Lemeshow test may fail [21], [22]. This can be observed for $t \geq 16$ as results on both sets don't match : the three models perform better on the test set due to standardization default. Regarding the discrimination, with large t there are not enough events to distinguish them, and this leads to poor scores.

The two models using a negative binomial distribution have fair results, in particular from a calibration perspective. For a given time-point t, the area under the ROC curve is approximately the same for the three models. Calibration of the single-part model is better than calibration of the two-part model according to the standardized Hosmer-Lemeshow test. It can be explained by the fact that Model 1 is trained on the full train set while the second part of Model 2 is trained

only on positive data. This constitutes a non-negligible loss of information.

Model 3 performs more poorly, especially for small time-point values where deviations between predictions and observations can be observed graphically.

VIII. Conclusions

This paper presents a novel approach to model train arrival delay risk at a major station with single and two-part generalized linear models. Prediction can be done a few days before operations in order to integrate the estimated probabilities in the scheduling process.

However, goodness-of-fit is more difficult to assess for probabilistic predictions as they cannot be compared with observations directly. A validation methodology is proposed based on their calibration and discrimination ability.

Both models using a negative binomial distribution achieve good calibration and acceptable discrimination results. These results may be improved by adding new predictors or modify their shape and encoding. Indeed, performance of any method will be bounded by the quality of the predictors. Another option would be to clean data and delete outliers.

Future researches will focus at first on using this methodology to other train types of the Montparnasse station and then on integrating this predictions in operations research models for main stations management.

Acknowledgement

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

The authors thank the OpenGOV team at DGEX Solutions-SNCF Reseau for their help on the resolution of the platforming problem and the encountered robustness issues.

References

[1] M. Abril, F. Barber, L. Ingolotti, M. Salido, P. Tormos, and A. Lova, "An assessment of railway capacity," *Transportation Research Part E: Logistics and Transportation Review*, vol. 44, no. 5, pp. 774–806, 2008.

[2] A. Stathopoulos and T. Tsekeris, "Methodology for processing archived its data for reliability analysis in urban networks," in *IEE Proceedings-Intelligent Transport Systems*, vol. 153, no. 1. IET, 2006, pp. 105–112.

[3] N. Marković, S. Milinković, K. S. Tikhonov, and P. Schonfeld, "Analyzing passenger train arrival delays with support vector regression," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 251–262, 2015.

[4] J. Peters, B. Emig, M. Jung, and S. Schmidt, "Prediction of delays in public transportation using neural networks," in *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, vol. 2. IEEE, 2005, pp. 92–97.

[5] S. Pongnumkul, T. Pechprasarn, N. Kunaseth, and K. Chaipah, "Improving arrival time prediction of thailand's passenger trains using historical travel times," in *Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on*. IEEE, 2014, pp. 307–312.

[6] P. Kecman and R. M. Goverde, "Predictive modelling of running and dwell times in railway traffic," *Public Transport*, vol. 7, no. 3, pp. 295–319, 2015.

[7] L. Oneto, E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, and D. Anguita, "Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 10, pp. 2754–2767, 2017.

[8] F. Cerreto, B. F. Nielsen, O. A. Nielsen, and S. S. Harrod, "Application of data clustering to railway delay pattern recognition," *Journal of Advanced Transportation*, vol. 2018, 2018.

[9] C.-W. Palmqvist, N. O. Olsson, L. W. Hiselius, and L. Winslott, "Punctuality problems from the perspective of timetable planners in sweden," in *Proceedings of the 20th EURO Working Group on Transportation Meeting, EWGT 2017*, 2017.

[10] M. Carey and S. Carville, "Testing schedule performance and reliability for train stations," *Journal of the Operational Research Society*, vol. 51, no. 6, pp. 666–682, 2000.

[11] A. Landex, A. H. Kaas, and O. A. Nielsen, *Methods to estimate railway capacity and passenger delays*. Technical University of Denmark (DTU), 2008.

[12] A. Agresti, *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[13] R. M. Goverde, I. Hansen, G. Hooghiemstra, and H. Lopuhaä, "Delay distributions in railway stations," *Proceedings of the 9th WCTR*, 2001.

[14] J. Yuan, "Stochastic modelling of train delays and delay propagation in stations," Ph.D. dissertation, TU Delft, Delft University of Technology, 2006.

[15] Y. Min and A. Agresti, "Modeling nonnegative data with clumping at zero: a survey," *Journal of the Iranian Statistical Society*, vol. 1, no. 1, pp. 7–33, 2002.

[16] R. D'agostino and B.-H. Nam, "Evaluation of the performance of survival analysis models: discrimination and calibration measures," *Handbook of statistics*, vol. 23, pp. 1–25, 2003.

[17] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[18] D. W. Hosmer and S. Lemesbow, "Goodness of fit tests for the multiple logistic regression model," *Communications in statistics-Theory and Methods*, vol. 9, no. 10, pp. 1043–1069, 1980.

[19] D. M. Stasinopoulos, R. A. Rigby *et al.*, "Generalized additive models for location scale and shape (gamlss) in r," *Journal of Statistical Software*, vol. 23, no. 7, pp. 1–46, 2007.

[20] D. Stasinopoulos, R. Rigby, and C. Akantziliotou, "Instructions on how to use the gamlss package in r," *Accompanying documentation in the current GAMLSS help files,(see also http://www. gamlss. org/)*, 2006.

[21] P. Paul, M. L. Pennell, and S. Lemeshow, "Standardizing the power of the hosmer–lemeshow goodness of fit test in large data sets," *Statistics in medicine*, vol. 32, no. 1, pp. 67–80, 2013.

[22] D. W. Hosmer, S. Lemeshow, and J. Klar, "Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small," *Biometrical Journal*, vol. 30, no. 8, pp. 911–924, 1988.