Evaluating stochastic train process time distribution models on the basis of empirical detection data

J. Yuan, R. M. P. Goverde & I. A. Hansen
Faculty of Civil Engineering and Geosciences,
Delft University of Technology, The Netherlands

Abstract

This paper evaluates several commonly applied probability distribution models for stochastic train process times based on empirical data recorded in a Dutch railway station, The Hague Holland Spoor. An initial guess of model parameters is obtained by the Maximum Likelihood Estimator (MLE). An iterative procedure is then followed, in which large delays are omitted one by one and the distribution parameters are estimated correspondingly using the MLE method. The parameter estimation is improved by minimizing the Kolmogorov-Smirnov (K-S) statistic where of course the empirical distribution is still based on the complete data set. A local search is finally performed in the neighbourhood of the improved model parameters to further optimize the estimation. To evaluate the distribution models, we compare the K-S statistic among the fitted distributions with optimized parameters using the one-sample K-S goodness-of-fit test at a commonly adopted significance level of $\alpha = 0.05$. It has been found that the log-normal distribution can be generally considered as the best approximate model among the candidate distributions for both the arrival times of trains at the platform and at the approach signal of the station. The Weibull distribution can generally be considered as the best approximate distribution model for non-negative arrival delays, departure delays and the free dwell times of late arriving trains. The shape parameter of the fitted distribution is generally smaller than 1.0 in the first two cases, whereas it is always larger than 1.0 in the last case. These distribution evaluation results for train process times can be used for accurately predicting the propagation of train delays and supporting timetable design and rescheduling particularly in case of lack of empirical data.

Keywords: train delays, running and dwell times, track occupancy times, statistical distribution, the K-S test.

1 Introduction

Modelling the distribution of train process times is an important research topic. Arrival and departure delay distributions reflect the punctuality level of trains at stations. Based on the distribution of input delays of trains at the boundary of a railway network and the distribution of primary delays within this network, the distribution of knock-on delays and that of the resulting exit delays can be estimated, which supports timetable design and operations management [10].

Train process time distributions are often assumed based on experiences from real operations and limited literature [2], [3], [6], [8], [9] exists with respect to statistical inference of the distributions using empirical data observations. Track occupation and release records show the total delays of trains and may include knock-on delays. Therefore, data filtering is necessary to fit the distribution of primary delays on the basis of train detection data [9]. To the best of our knowledge, there is no publication that evaluates the conditional train running and dwell and track occupancy time distributions in the case of different aspects of relevant block signals, which can be used for estimating knock-on delays more accurately [10].

This paper evaluates several commonly applied distribution models for stochastic train process times on the basis of empirical traffic data recorded in a Dutch railway station The Hague Holland Spoor (The Hague HS). The evaluation of distribution models is performed not only for the arrival times (delays), non-negative arrival delays and departure delays of trains at the station, but also for the arrival times of trains at the boundary of the local railway network and the train running, dwell and track junction occupancy times within the local network. This paper is structured as follows. Section 2 outlines the distribution models to be evaluated and the evaluation method. The results of the distribution evaluation for the process times of trains are then discussed in Section 3. Finally, the main conclusions are drawn in Section 4.

2 Distribution models and the evaluation method

A variety of theoretical probability distributions, including the normal, uniform, exponential, gamma, beta, Weibull, and log-normal distributions [5], have been adopted in the literature to model the stochasticity of train process times. Given a continuous distribution model, the parameters are estimated on the basis of empirical data using e.g. the maximum likelihood method. The resulting fit can be tested using the Kolmogorov-Smirnov (K-S) goodness-of-fit test. This test is based on the K-S statistic, defined as the maximum absolute difference between the empirical and fitted cumulative distribution function [5]. However, the estimated parameters are generally sensitive to outliers in the data set. We therefore used an iterative parameter estimation method. An initial guess of the model parameters is obtained using the Maximum Likelihood Estimator (MLE) of the complete data set. Next, the large delays in the original data set are omitted iteratively one by one estimating the distribution parameters correspondingly using the MLE method. In each iteration, we compute the K-S statistic where of course the empirical distribution is still based on the complete data set. The iterative procedure terminates if the K-S statistic cannot be decreased any more. After the iterative procedure, we apply a local search in the neighbourhood of the parameter estimate to further optimize the parameter estimation by minimizing the K-S statistic. It should be mentioned that the location parameter of each distribution model except for the normal distribution was taken as the minimum value of the empirical data observations.

To evaluate the candidate distributions, we compare the K-S statistic among the fitted distributions with optimized parameters using the one-sample K-S test [5] at a commonly adopted significance level $\alpha = 0.05$. To visualize the quality of distribution fitting for the process times of trains, we compare the fitted distribution density curve with the kernel estimate and empirical histogram [7] and apply the distribution differences plot [1] for the fitted distribution and the empirical one.

3 Evaluation results

The distribution of train process times may depend on the types and routes of trains. We hence evaluate the distribution models for the process times of trains per train series in both the southbound and northbound directions at The Hague HS railway station.

3.1 Arrival times

The modelling of train arrival time distribution is a prerequisite for predicting the propagation of train delays at stations. To incorporate the impact of the knock-on delays caused by route conflicts in a delay propagation model, we need to distinguish the arrival times of trains at the station platform from that at the approach signal of the station. Early arriving trains are often considered as punctual trains in some delay propagation models [4], where the distribution of non-negative arrival delays is of the main concern.

It has been found that the location-shifted log-normal distribution is the best approximate model among the candidate distributions in 9 and 11 of the 14 considered cases for both the arrival times of trains at the approach signal of the station and at the platform track, respectively. This distribution model has not been rejected by the K-S test in 9 and 10 of the 14 cases for both the arrival times. The optimized parameters and the K-S test results for the log-normal distribution are given in Table 1, where μ and σ represent the mean and standard deviation of the underlying normal distribution and p denotes the p-value [5] of the K-S test.

Figure 1: shows the optimized log-normal distribution density curve, kernel estimate and empirical histogram for the arrival times of the northbound intercity train series IC2100N at the approach signal of the station. The corresponding distribution differences plot is shown in Figure 2:, where the two dotted horizontal lines represent the critical error bounds for the K-S test. In both figures, the reference time is defined at the scheduled arrival time of the studied

train series at the station. The optimized log-normal fit matches well the empirical data. In this case, the distribution differences plot does not cross the error bounds, therefore the location-shifted log-normal distribution has not been rejected by the K-S test.

Table 1: Optimized parameters and the K-S test results of the log-normal fit for the arrival times of trains in The Hague HS.

Train	At the approach signal				At the platform track			
Series	μ	σ	Best	*p*	μ	σ	Best	*p*
IR2200S	4.6	0.6	×	0.03	4.8	0.6	√	0.16
IC2100S	4.3	0.7	√	0.61	4.5	0.7	√	0.38
IC2400S	4.2	0.8	√	0.25	4.4	0.9	√	0.42
INT600S	4.9	0.7	×	0.18	5.0	0.5	√	0.20
HST9300S	4.8	0.9	×	0.01	4.8	0.9	×	0.00
AR5000N	4.7	0.4	√	0.00	4.7	0.5	√	0.00
AR5100N	4.5	0.5	√	0.00	4.6	0.5	√	0.00
IR2200N	4.6	0.4	×	0.00	4.5	0.5	×	0.00
IC1900N	4.9	0.6	×	0.15	5.1	0.5	×	0.16
IC2100N	5.0	0.7	√	0.55	5.0	0.7	√	0.14
IC2400N	4.8	0.6	√	0.31	4.9	0.6	√	0.79
IC2500N	4.8	0.6	√	0.13	4.9	0.5	√	0.15
INT600N	4.8	0.7	√	0.26	4.9	0.6	√	0.22
HST9300N	4.6	0.7	√	0.38	4.6	0.8	√	0.30

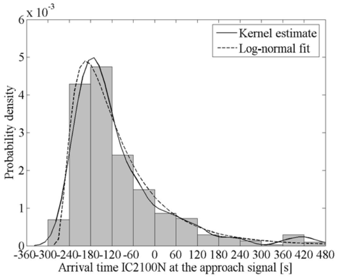

Figure 1: Log-normal fit, kernel estimate and histogram for the arrival times of IC2100N at the station approach signal.

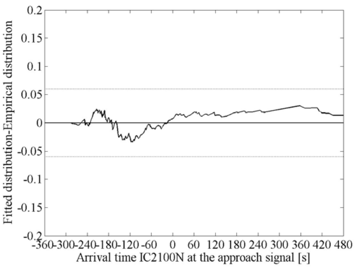

Figure 2: Distribution differences plot for the log-normal fit and the arrival times of IC2100N at the station approach signal.

We have also evaluated the candidate distributions for non-negative arrival delays of trains at the platform track. The Weibull distribution is the best fit among the candidate distributions in 7 of the total 18 studied cases. This distribution model has not been rejected by the K-S test in 17 of the 18 cases. Since the exponential distribution has been widely used to model the stochasticity of non-negative arrival delays [2], [6], [9], we visualize the goodness-of-fit of both the Weibull and exponential distributions in Figure 3: and Figure 4: for the northbound intercity train series IC2100N. The non-negative arrival delays fit to the Weibull distribution with a shape parameter of 0.8 overall better than the exponential distribution especially at the range of larger delays. Both the distributions have not been rejected by the K-S test in this case.

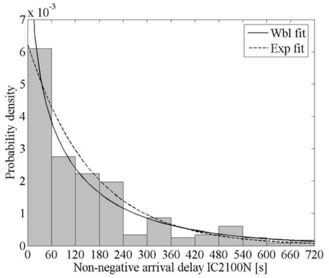

Figure 3: Fitted density curves and histogram of non-negative arrival delays of IC2100N.

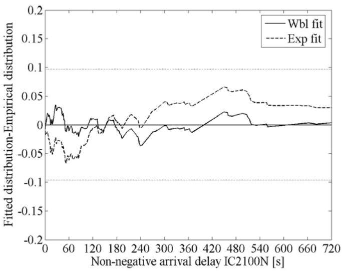

Figure 4: Distribution differences plots for non-negative arrival delays of IC2100N.

In conclusion, the log-normal distribution can be generally considered as the best model among the candidate distributions for both the arrival times of trains at the platform and at the approach signal of the station. The Weibull distribution matches well non-negative arrival delays. For simplicity, we may use the exponential distribution, which is a special type of the Weibull distribution, to be as an approximate distribution model for non-negative arrival delays if the density is decreasing.

3.2 Departure delays

Departure delays are non-negative since trains are not allowed to depart from the station earlier than the scheduled departure time. The distribution of departure delays can be used to predict the distribution of outbound track release times and the distribution of train arrival times at the following stations.

It has been found that the Weibull distribution is the best approximate model among the candidate distributions for the departure delays of trains in 11 of the total 18 studied cases. The exponential distribution is the best approximate model in 2 of the 18 cases. In addition, both the distributions have not been rejected by the K-S test in 10 of the total 18 cases. Figure 5: and Figure 6: visualize the goodness-of-fit of both the Weibull and exponential distributions for the departure delays of the southbound intercity train series IC2400S. Early arriving trains usually do not depart very late and some trains may arrive at and depart from the station very late, which results in a very steep histogram of the departure delays. The Weibull fit with a shape parameter of 0.8 matches the delays overall better than the exponential fit. In this case, the former distribution has not been rejected while the latter distribution has been rejected by the K-S test. Thus, we can generally consider the Weibull distribution to be as the best approximate model among the candidate distributions for departure delays. Just like for non-negative arrival delays, the exponential distribution can be considered as an approximate distribution model for departure delays if the density is decreasing.

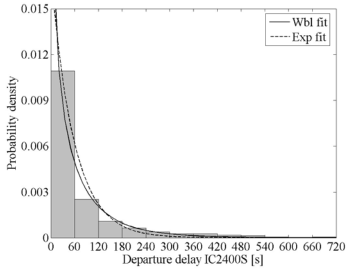

Figure 5: Fitted density curves and histogram of departure delays of IC2400S.

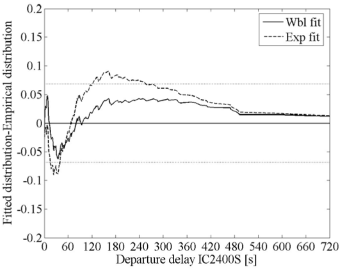

Figure 6: Distribution differences plots for departure delays of IC2400S.

3.3 Dwell times

The dwell times of trains are the difference between the arrival and departure times. Early arriving trains generally have much longer dwell times and they are not of our main concern. To estimate the knock-on delays and departure delays of trains at stations, it is critical to obtain the distribution of the free dwell times for late arriving trains [10]. The free dwell time of a train is defined as the necessary dwell time for passenger alighting and boarding in the absence of hindrance from other trains.

We fit the free dwell time distribution for late arriving trains per train series in one direction using the dwell time observations of the trains that satisfy

$$A_i > a_i \text{ and } C_i \leq A_i + 30$$

where, A_i and a_i denote the actual and scheduled arrival time of train i at the platform track and C_i represents the clearance time of the outbound route of this train. The unit of these times is in seconds. Late arriving trains are selected by the former inequality and the latter inequality ensures that the chosen trains are not hindered by other trains at the station after a minimal dwell time of 30 s.

It has been found that the Weibull distribution is the best approximate model among the candidate distributions for the free dwell times of late arriving trains in 16 of the total 18 studied cases. In addition, this distribution model has not been rejected by the K-S test in all the cases. Figure 7: and Figure 8: visualize the goodness-of-fit of the Weibull distribution model with a shape parameter of 1.9 in case of the northbound interregional train series IR2200N. The fitted distribution matches well the kernel estimate for the empirical data and it has not been rejected by the K-S test. In conclusion, the Weibull distribution with a shape parameter larger than 1.0 is the best approximate model among the candidate distributions for the free dwell times of late arriving trains.

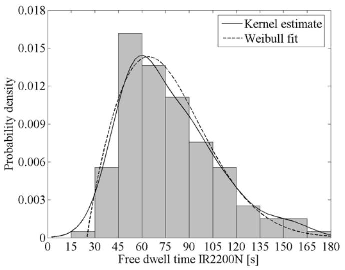

Figure 7: Weibll fit, kernel estimate and histogram for the free dwell times of late arriving trains of IR2200N.

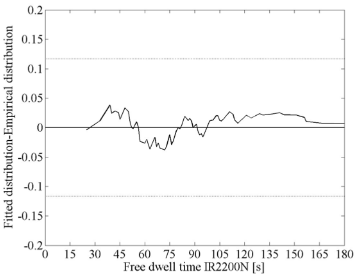

Figure 8: Distribution differences plot for the Weibull fit and the free dwell times of late arriving trains of IR2200N.

3.4 Running times and track occupancy times

The distribution of train running times and that of track occupancy times are required to estimate the propagation of train delays in a railway network. In this paper, we focus on statistical distribution of the running times of trains on the preceding block of The Hague HS station and that of the occupancy times of adjacent junctions around this station.

In case of an approaching train, if the inbound route is released earlier than the time of the train arriving at sight distance of the approach signal, the train approaches the station at the free running speed. Otherwise, the train is hindered and has to decelerate and even stop on the preceding block of the station. To accurately estimate the knock-on delays caused by route conflicts in a station area, it is necessary to investigate the conditional distributions of inbound train running and track occupancy times in the case of different aspects of the approach signal and home signal of the station. For a departing train, if it is hindered due to outbound route conflicts, it dwells at the station for a longer time, but running on the next track sections will not be hindered again. In this case, the conditional distributions are not applicable.

To model the conditional distributions of inbound train running and track occupancy times based on a statistical analysis of the empirical data, the first step is to classify the data observations. By comparing the arrival time of each train at the approach signal to the clearance time of the inbound route, we have extracted a data set suited for fitting the free train running and track occupancy time distributions in each studied case. A hindered approaching train may pass the home signal with reduced speed without a stop or with acceleration speed after a stop in front of this signal. Since the standstill of a train on track is not recorded, we cannot directly identify whether or not a hindered train stops before the home signal based on track occupancy and release records. Adopting the k-means routine within the statistical analysis tool S-Plus [7], we have split the data sample of hindered trains for each studied train series into two separate parts which correspond approximately to the two cases mentioned in the above. However, for the hindered trains that stop before the home signal, it is still unknown when these trains stop. Therefore, we have lack of the running times of these trains on the preceding block of the station.

For the free running times of trains on the preceding block of the station, most of the candidate distributions have been rejected by the K-S test. Both the Weibull and normal distributions have not been rejected by the K-S test in 2 of the total 13 studied cases. In addition, each of these distributions is the best approximate model among the candidate distributions in 5 of the 13 cases. For inbound junction occupancy times by the free passing trains, the Weibull distribution is the best approximate distribution model in 3 of the total 4 considered cases and has not been rejected by the K-S test. The normal and Weibull distribution is the best fit among the candidate distributions for outbound track occupancy times in 2 and 1 of the total 3 considered cases, respectively. In addition, both the distributions have not been rejected by the K-S test in 2 of the 3 cases. The goodness-of-fit of the Weibull and normal distributions for the above-mentioned train process times is shown in Figure 9: and Figure 10:, respectively.

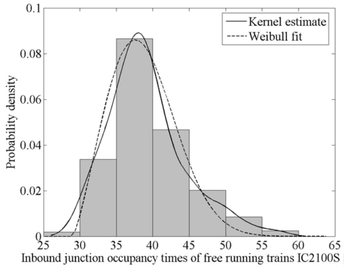

Figure 9: Weibull fit, kernel estimate and histogram for inbound junction occupancy times of free running trains IC2100S.

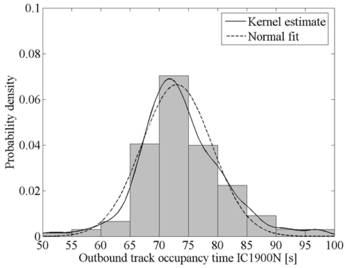

Figure 10: Normal fit, kernel estimate and histogram for outbound junction occupancy times of IC1900N.

In the case of the hindered trains that do not stop before the home signal, the Weibull and normal distribution is the best approximate distribution model for the running times on the preceding block of the station in 2 and 1 of the total 6 considered cases, respectively. In addition, both of the distributions have not been rejected by the K-S test in the 6 cases. For inbound junction occupancy times by the hindered trains, the normal and Weibull distribution is the best approximate distribution model in 2 and 1 of the 5 considered cases and these distributions have not been rejected by the K-S test in the 5 cases. In the case of the hindered trains that stop before the home signal, the inbound junction occupancy times fit best to the normal distribution in the two considered cases and this distribution has not been rejected by the K-S test in one of the two cases. The goodness-of-fit of the Weibull and normal distributions for the above-mentioned train process times is given in Figure 11: and Figure 12:, respectively.

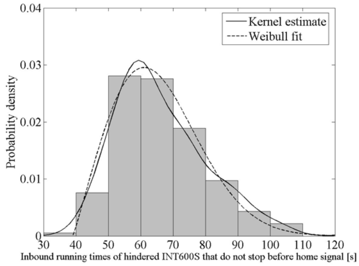

Figure 11: Weibull fit, kernel estimate and histogram for inbound running times of hindered INT600S that do not stop before the home signal.

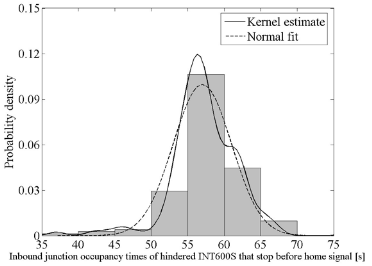

Figure 12: Normal fit, kernel estimate and histogram for inbound junction occupancy times of hindered INT600S that stop before the home signal.

In conclusion, it is difficult to find a good distribution for the conditional train running and track junction occupancy times in the case of different aspects of relevant block signals. This might be because of the big variation of train speed on the short track sections in the complicated station and junction area.

4 Conclusions

We have compared several commonly applied distribution models for train process times on the basis of empirical train detection data recorded at a Dutch railway station The Hague HS. It has been found that the log-normal distribution can be generally considered as the best approximate model among the candidate distributions for both the arrival times of trains at the platform and at the approach signal of the station. The Weibull distribution can generally be considered as the best approximate distribution model for non-negative arrival delays, departure delays and the free dwell times of late arriving trains. The shape parameter of the fitted distribution is generally smaller than 1.0 in the first two cases, whereas the shape parameter is always larger than 1.0 in the last case. For simplicity, the exponential distribution can be used as an approximate distribution model for non-negative arrival delays and departure delays if the density is decreasing.

Acknowledgement

This publication is a result of the research programme Towards Reliable Mobility, carried out within the Transport Research Centre of Delft University of Technology.

References

[1] Averill, M.L. & Kelton, W.D., Simulation Modeling and Analysis, McGraw-Hill, 2000.

[2] Goverde, R.M.P., Hooghiemstra, G. & Lopuhaä, H.P., *Statistical Analysis of Train Traffic: The Eindhoven Case,* DUP Science, Delft, 2001.

[3] Hermann, U., *Untersuchung zur Verspätungsentwicklung von Fernreisezügen auf der Datengrundlage der Rechnerunterstützten Zugüberwachung Frankfurt am Main,* PhD thesis, Technische Hochschule Darmstadt, 1996.

[4] Radtke, A. & Hauptmann, D., Automated planning of timetables in large railway networks using a microscopic data basis and railway simulation techniques, In: Allan, J. et al. (eds.), *Computers in Railways IX*, pp. 615-625, WIT Press, Southampton, 2004.

[5] Ross, S. M., *Introduction to Probability and Statistics for Engineers and Scientists*, Elsevier, 2004.

[6] Schwanhäusser, W., *Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn*, PhD thesis, RWTH Aachen, 1974.

[7] S-Plus, *S-Plus 2000 Guide to Statistics*, Vol. 2, Data Analysis Products Division, Mathsoft, Seattle, 1999.

[8] Steckel, J., *Strategische Optionen für die Zufällige Fahrzeit im Eisenbahnbetrieb,* PhD thesis, Hochschule für Verkehrswesen 'Friedrich List' Dresden, 1991.

[9] Wendler, E. & Naehrig, M., Statistische auswertung von verspätungsdaten, *Eisenbahningenieurkalender EIK*, pp. 321-331, 2004.

[10] Yuan, J. & Hansen, I.A., Optimizing Capacity Utilization of Stations by Forecasting Knock-On Train Delays, In: Hansen, I.A. et al. (eds.), *Proceedings of 1st International Seminar on Railway Operations Modelling and Analysis*, Delft, 8-10 June, 2005.