Danish Journal of Transportation Research

Distribution Fitting for Very Large Railway Delay Data Sets with Discrete Values

Steven Harrod[a],*, Georgios Pournaras[c], Bo Friis Nielsen[b]

[a]Department of Technology, Management, and Economics,
Technical University of Denmark
[b]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[c]Ansaldo STS - a Hitachi company group

*Corresponding author: stehar@dtu.dk

Article info

Article history:
Received 17/08/2018
Received in revised form
26/06/2019
Accepted 19/09/2019

Keywords:
Big data, railway delays,
delays distribution

Abstract

Modern railway signal systems allow the collection of very large data sets (more than a thousand values). These data sets are often rounded by the signal technology, so that the values are effectively discrete. This paper reviews other literature on fitting distributions to large data sets, and then shares the experience of distribution fitting to a large data set from the Danish railways.

1. Introduction

Modern railway signal and control systems are capable of collecting and storing very large quantities of train operation data. This data is increasingly available to researchers and management staff, and can be studied on common desktop computers. However, the large volume of data reveals new patterns of behavior that in some cases contradict prior findings. This paper reviews some prior literature on the subject of probability distribution fitting to operations data, and shares some learned best practices from analysis of a Danish railway line with high traffic density.

The most common analysis activity is the study of punctuality data, either to assess train service performance for managerial or contractual activities, or to understand the train performance so that services may be modified and improved in the future. Very frequently in the literature, the measures of interest are the deviations (and usually delays) from plan or timetable of train arrivals, train departures, and the time spent idle at the platform (dwell time). The statistics and distribution of this data are frequently important to management, as they often form the basis of contractually agreed customer service levels. The data should also be understood in detail to understand the impact of new or revised customer service commitments.

Statistical analysis of and distribution fitting to this data is also important to operations planners. The analysis of the data can reveal patterns of operation and potential control variables [Cerreto et al., 2018]. Most commonly, theoretical distributions are fitted to the data so that train operations may be simulated or modeled with queuing theory [Welch and Gussow, 1986, Hallowell and Harker, 1998, White, 2005, Lindfeldt, 2015, Jovanovic et al., 2017]. Vromans [2005] demonstrates how different assumptions of delay distribution can significantly change the simulation output response, due to the nonlinear response of railway systems to delay events. It is important that these distributions are correctly fitted to the observed processes, and that is the primary subject of this paper.

Section 2 offers a summary of goodness of fit theory and lessons learned from other areas of science. Section 3 reviews some published analysis of punctuality data from other railways. Finally, lessons learned from fitting distributions to data from the Danish Kystbane are discussed in Section 4 and concluding remarks and future research interests are offered in Section 5.

2. Methods of Data Fitting and Lessons from Other Science Fields

The most common method for fitting a distribution to a sample data set is the maximum likelihood procedure. Commercial software (in this paper, SAS 9.4 M4) estimates the parameters for a range of common distributions based upon the data set entered. For each distribution (e.g. normal, exponential, uniform, etc.), parameters are proposed that "maximize" the probability that the data set was sampled from the proposed distribution. The result is a list of candidate theoretical distributions that could represent the data, but there is no guarantee that any of the candidates are appropriate. That is, the maximum likelihood method proposes parameters that best fit that distribution to the data, but the resulting distribution may still be a very poor fit to the data.

The next step is to judge which, if any, of these distributions are in fact a "good" fit to the data set. The most common method of judging this "goodness of fit" is the one sample Kolmogorov-Smirnov test [Massey, 1951]. In most commercial software, the test reports a "p-value" that represents the probability that a more extreme data set than the one under consideration could have been generated with the distribution under consideration. The smaller the p-value, the less likely that the data set under consideration could have been drawn from the reference distribution. In this case, small p-values imply that the sample data set is behaving significantly different than the controlling distribution, and that it is very unlikely that the proposed distribution generated the data studied. Very often $p \leq ,05$ is selected as a key measure for rejecting the goodness of fit.

A direct challenge to distribution fitting in large data sets is that the Kolmogorov- Smirnov test statistics are directly a function of the number of data points (size of sample, n). As the data set approaches an infinite number of samples, the test statistic becomes infinitely strict. This can create challenges to fitting a theoretical distribution if, for example, the data contains biased noise, or other patterns of behavior [Johnson and Wichern, 2007].

Browne and Cudeck [1992] discusses at length the question of what defines "good fit". They note that often when data is dirty or originating from a real process, no distribution will truly fit the data. Paradoxically, when the data sample is small, theoretical distributions will appear to have a good fit. When the data sample from the same process is larger, the distributions will often fail the goodness of fit test. Browne and Cudeck states, "Statistical goodness-of-fit tests are often more a reflection on the size of the sample than on the adequacy of the model" and "Model selection has to be a subjective process involving the use of judgment." Browne and Cudeck reviews the various statistical methods, and then propose that the hypothesis test should not seek to achieve a perfect error sum of zero, but should seek some subjective allowable error such as 5%. They call this the "null hypothesis of close fit".

The exponential distribution is a fundamental distribution in many applications such as queuing theory and simulation. In the absence of sufficient data, it is often recommended as a default modeling assumption [Law and Kelton, 2000, Harrod and Kanet, 2013]. Bolotin [1994] examines the fit of the exponential

distribution to telephone circuit holding times. The research considers data sets of greater than n = 1.000, and finds that the ordinary exponential distribution is a poor fit to the behavior. One discovery presented is that the telephone data fails to adhere to the "memoryless" property of the exponential distribution. In short, the distribution of future events in the exponential distribution should be independent of the completed events, but in the telephone circuit data there is a very clear growth in the mean remaining circuit hold time as elapsed circuit time increases. In Section 4, this will be shown to be the case also for railway delay data.

Bolotin [1994] demonstrates that a log normal distribution (to base 10) is a much better fit to the telephone circuit data. The research then further demonstrates that aggregate data can clearly be decomposed into identifiable groups based on specific call types (voice, fax machine, data), and then the aggregate distribution is best represented by a mixture distribution of log normals. The research further hypothesizes that the log normal distribution of the call time originates from the human perception of time, and a psychophysical law called "Weber's Law". The principal of this theory, is that in order for stimulus to be registered by the psychi, it must increment in proportion to the stimulus that is already present.

3. Some Prior Studies on Railway Punctuality Data

Many prior papers consider the distribution fitting and analysis of railway punctuality data. A sample of prior studies are reviewed here in chronological order. Goverde [2005] offers what may be the earliest extensive study of railway punctuality (see also Goverde [2001]). This study examines a data set of 16 trains per hour from one week of September, 1997 at Eindhoven railway station. The data is segregated by train service, and there are 13 listed services with an average sample size of n = 93 in arrival data and n = 103 in departure data. Data is further grouped as arrival delay, departure delay, and dwell times, and evaluated separately. Exponential distributions fit 10 out of 13 train services with $p \geq ,05$ from the Kolmogorov-Smirnov test. Seven of the train services are through services, and normal distributions fit all of their dwell times with the smallest p-value recorded at 0,31.

This same data set forms the basis of further selective studies such as Goverde and Hansen [2001], which finds that recorded delays show much higher standard deviation than mean. The data demonstrates non-independent behavior such as dwell times of late arriving trains exceeding scheduled dwell time, which is counter intuitive to timetable design expectations. Also discussed in detail is how data collected from the signal system does not represent actual platform times and must be adjusted and cleaned for many factors. The Netherlands Railways tool for this activity is called TNV-Prepare/Filter. In most cases, the last measurement point before a station is 1 km out, and the departing measurement point is located at the platform exit signal. The stopping point of trains may also vary by train length and passenger access point location on the platform.

Yuan and Hansen [2002] and Nie and Hansen [2005] should be read together, as they discuss different aspects of the same data set. Both papers consider data from train movements in and around Hague Holland Spoor and Hague Central stations in September, 1999, and both papers use the data management methods of Goverde and Hansen. Yuan and Hansen examines a data set of arrivals and departures at Hague Holland Spoor, consisting of 24 arrivals and departures per hour, 450 trains per day, and 10.000 data points. The study notes that the data demonstrates that the longer the train route, the higher the deviation. Fitting of distributions is by maximum likelihood method and single sample Kolmogorov-Smirnov test. Trains are divided into 24 classifications by route and train type, and distributions fitted to these groups, approximately n = 416 per group. For 18 of these groups, late arrivals fit to an exponential distribution and excess dwell times fit to a normal distribution.

Nie and Hansen [2005] performs a detailed micro analysis of the traffic between stations Hague Holland Spoor and Hague Central, a distance of 1,65 km. The data (September 1999) consists of about 8 trains per hour, or 4.320 data points. The study finds that a normal distribution fits the running time, but neither the departure nor arrival delays could fit a distribution satisfying the Kolgomorov-Smirnov test at $\alpha = ,05$. It is

interesting to note that this is essentially the same data set under analysis in the same research group, but what is different is that the sample size in Nie and Hansen is approximately ten times larger than that in Yuan and Hansen.

Yang et al. [2017] fits distributions to the delay data of the Guangzhou Railway Corporation high speed line. The data set has 11.452 delay events, including delays causes, over 8 months in 2015. Yang et al. does not clearly state whether the data is arrival or departure delays. Distribution fitting is by maximum likelihood and Kolmogorov-Smirnov, with some additional investigation of skewness and kurtosis by Cullen and Frey graphs. The primary finding is that a lognormal distribution is frequently appropriate. When the data is segregated by delay cause (and sample size is smaller), the p value for goodness of fit is significantly higher. The p value for goodness of fit for the aggregate data distribution is ,06. Finally, Wen et al. [2017] examines a data set of only primary delays on the high speed railway between Wuhan and Guangzhou, China. The data set consists of 1.249 records over ten months in 2015. Lognormal distributions are proposed as best fitting the data.

4. Experience Gained from Study of the Danish Kystbane

As part of the IPTOP project [Nielsen, 2017], over five years of operating data is collected in a single database for analysis. In this section, distributions are fitted to departure delays from a small portion of this data, 75.244 records (!). The data comes from the "Kystbane", the coastal railway running north from Copenhagen, from weekdays in the period September through November, 2014, inclusive. Only northbound traffic passenger traffic is considered. Freight traffic is negligible and typically only at night, and is excluded. Three services are operated: ØP trains running in "Øresund" service from Sweden (26.001 data points), ØK trains running in local coastal service only (44.679 data points), and ØD trains running in peak demand periods only (4.564 data points). The data set covers eleven stations, from Østerport to Snekkersten, with an average distance between stations of 3,98 km. The three train services have different stopping patterns. Similar to the Netherlands railways, this data is collected from the signal system, and is adjusted to correctly identify times at the platform in stations [Richter et al., 2013].

The first experiment is to attempt to fit a distribution to the complete data set. The normal, lognormal, and exponential distributions are fitted to the data, and none of them offer a Kolmogorov-Smirnov (KS) p-value greater than 0,001. All three are bad fits to the data. The best fitting, by judgment of the probability plot, is the exponential shown in Figure 1, which displays the histogram, the fitted exponential distribution, and the probability plot in inset. The deviation in the probability plots for the normal and lognormal distributions (not shown) is quite extreme.

Note that the data contains some early departure records. If the data is limited to only late departures (deviation greater than zero, n = 55.585), there are still no acceptable model fits, and the probability plots are even worse in their display. However, for the rest of this analysis, only late departures will be studied, as they are typically of greater interest when judging timetable performance. If the sample size is limited, the lognormal distribution becomes acceptable in its fit, which follows the theories and examples discussed in Section 2. For example, if random 0,5% samples of the data set ($n \approx 275$) are fitted, KS p-values for the lognormal goodness of fit ranged from a low of 0,073 to a high of 0,50. However, the probability plot still displays strong divergence at the upper quartile. That is, smaller portions of the larger data set pass the goodness of fit test, but as the sample size increases, it becomes difficult to evaluate the data set using standard methods.

Figure 2 repeats the memoryless property investigation of Bolotin [1994]. Under the memoryless property of the exponential distribution, the time that has passed should have no influence on the statistical distribution of the remaining values. To test this, 2.497 random samples from an exponential distribution with mean 2,5 minutes, the same as the Kystbane delay sample, are generated. Calculations are made at regular intervals, where the mean of the remaining delay is calculated. The remaining delay is represented by Equation (1), where α is the reference elapsed time, and δ is the original delay

$$\begin{array}{ll} \delta - \alpha & \delta > \alpha \\ null & otherwise \end{array} \quad (1)$$

The mean value of the positive remainders is then charted against the test level. The same values are calculated from the Kystbane data sample as well. As can be seen in the figure, the exponential sample exhibits a nearly constant remaining expectation, but the Kystbane data exhibits nearly linear growth. Bolotin described this as "the longer a [telephone] conversation goes, the longer it is likely to continue", and a similar analogy can be made about the Kystbane delays, the longer a train is delayed, the more severe a delay it is likely to be experiencing.

One consideration might be that the behavior is a function of the individual station, and that by limiting the analysis to a specific station, different (better) results might be obtained. Rungsted Kyst is an example of a candidate station for focused analysis. It is the ninth station northbound from Copenhagen. Only local service trains stop at this station, all other trains running through without stopping, and there are 4.494 late departure data points for Runsted Kyst. This limitation does not, however, lead to acceptable distribution fits. The results are nearly the same as for the whole data set analysis. When a random smaller (0,5%) sample of the data is fitted, in many cases lognormal and exponential distributions will appear acceptable.

From the preceding initial analysis, the exponential distribution is not a good representation of the late departure data. From the preliminary trials, and the cited literature, a lognormal distribution is a better candidate, if only the fit could be established for a large data sample. Bolotin [1994], previously discussed, also offers encouragement to consider mixed distributions. A mixed distribution is a blend of two or more distributions. It typically takes the form of $p(x) = w_1 f_1(x) + w_2 f_2(x) + \ldots$ where w is a proportion summing to 1. The distribution is then a collection of individual point values, each drawn from one of the functions $f_i(x)$, selected with probability w_i.

Mixed distributions are now accessible from commercial statistics software. SAS offers the FMM procedure. This analysis will use the R statistical software platform (version 3.5.1), and the "mixdist" package (version 0.5-5). Various trials were also conducted using the FMM procedure in SAS, and these guided the final conclusions here. One of the challenges in fitting mixed distribution models is that the result may be influenced by the initial model parameters supplied to the routine. The most critical input parameter is the number of expected component distributions. A common recommendation is to view the histogram of the data and try to count the number of component shapes.

The 55.585 records of late departure (delay greater than zero) were transformed to a log base 10. Examination of the histogram indicated at most two distribution components, and as expected, a mixture of normal distributions. Earlier trials in SAS where a mixed exponential distribution was fitted often showed three components. Another parameter set that must be entered is a set of recommended starting values for the mean and standard deviation of the components. Initially an effort was made to calculate these estimates by visually segregating the histogram in two overlapping halves, but in subsequent trials it was found that different starting values always lead to nearly the identical solution. The mixdist algorithm is quite efficient at finding the same solution quickly with nearly default values, for this class of problem.

Figure 3 presents the fitted distribution, which is $p(x) = 0{,}3792N(\mu = -0{,}3845, \sigma = 0{,}6478) + 0{,}6208N(0{,}2593, 0{,}4891)$. Plotted is a histogram, the sum of the two component distributions, and the two component distributions. Note this function is not equivalent to a single normal distribution. The two normal distributions are not blended 1:1. The combined distribution in Figure 3 is not symmetric about its mean.

Note the heavy point masses at the lower tail, which represent essentially on time trains that registered trivial delay values. Unfortunately, even though this is a good fit, the standard goodness of fit statistic offered by the software (in this case Chi-square, the only option in Mixdist), rejects the fit unconditionally. An examination of the probability plot in Figure 4 demonstrates an extremely good fit in the most

applicable data range, but extreme deviation at the lower tail. This is because the fitted distribution is returning a sample of delays very close to zero.

At this point a certain "executive decision" is made that the final distribution should not have this tail, and the distribution is modified to $p(x) = 0{,}3792max(-1{,}5, N(\mu = -0{,}3845, \sigma = 0{,}6478)) + 0{,}6208max(-1{,}5, N(0{,}2593, 0{,}4891))$. This truncates the data and concentrates the lower tail as a point mass of zero delay (not shown in the figure). Recall also that the earliness data has been discarded before the analysis began. In a modeling or simulation task, this non-delay should be included in the simulation logic or distribution behavior, typically as a fixed probability of early departure (although in many cases early departures are discouraged by management, and are a separate topic of discussion). With this modification, the probability plot is much improved as in Figure 5.

However, there is still that nagging problem of a lack of goodness of fit statistic. The goodness of fit is hindered by the extremely large sample size, and the small amount of discrete step function in the data, visible as the stair step pattern in Figures 4 and 5. This occurs because the current legacy signal system in Denmark rounds train movements to an accuracy of +/- ten seconds (this will change with the future ERTMS installation). The Netherlands railways record data to the second, and the Italian railways record data to increments of 30 seconds.

This discrete step pattern creates noise that disrupts the goodness of fit test. Further, it does not result in a proper discrete distribution, but a censored sample of the continuous delay distribution. Law and Kelton (2000) explain how discrete data can be difficult for both the Chi-square and Kolmogorov-Smirnov goodness of fits tests. In the case of the Kolmogorov-Smirnov test, critical values must be computed for each case with significant effort. For the Chi-square, the ideal test intervals should be equiprobable, but this is hard to accomplish in agreement with the intervals already defined by the discrete unit size.

An alternative goodness of fit test is the Wilcoxon rank sum test. Goverde (2001) utilizes this test frequently to judge whether the distribution of delay data is different between periods of day such as morning and afternoon. This may be performed in R statistical software with the function wilcox.test (paired = false). The process for conducting this test is presented in Figure 6. The application in Figure 6 is novel because the Wilcoxon test is utilized to assess goodness of fit. This test on the data in Figure 5 returned a satisfactory result of p = 0,678.

5. Conclusion

Modern data systems make very large data sets available to railway planners, but the statistical methods in common commercial statistics software are not designed to study these large data sets, and can frequently give misleading results. Some of the earlier published analyses of railway delays offer conclusions that may be incorrect because of the small sample size. Railway systems are frequently non-linear in their response to events, and small differences in the assumed probability distributions of events may lead to significant differences in analysis and simulation output.

Many earlier studies have recommended the exponential distribution as representative of railway processes, but closer examination of a large data set from the Danish railways shows that this is not a good default distribution. Examination of the data set supports a mixed distribution of lognormals, which is in agreement with research from the telephone industry. Care must be taken to remove or account for the large point mass of on-time or trivially late trains. The goodness of fit of these models and data sets can be supported with the Wilcoxon signed-sum test. All of these methods are accessible in advanced commercial software, but they are not the default options in software such as SAS, and no warning is provided to the user that alternative methods should be used. With some effort, the correct methods can be found in commercial software, but they must be invoked specifically.

The authors have identified the generalized linear model of exponential family distributions as another potential distribution model than can account for the full range of detail in the source data, such as location, train type, traffic, etc. Research in this area is ongoing.

Acknowledgement: This work was funded by a Dean Grant from the Technical University of Denmark (DTU) and by the Danish Innovation Fund through the IPTOP project (Integrated Public Transport Optimisation and Planning).

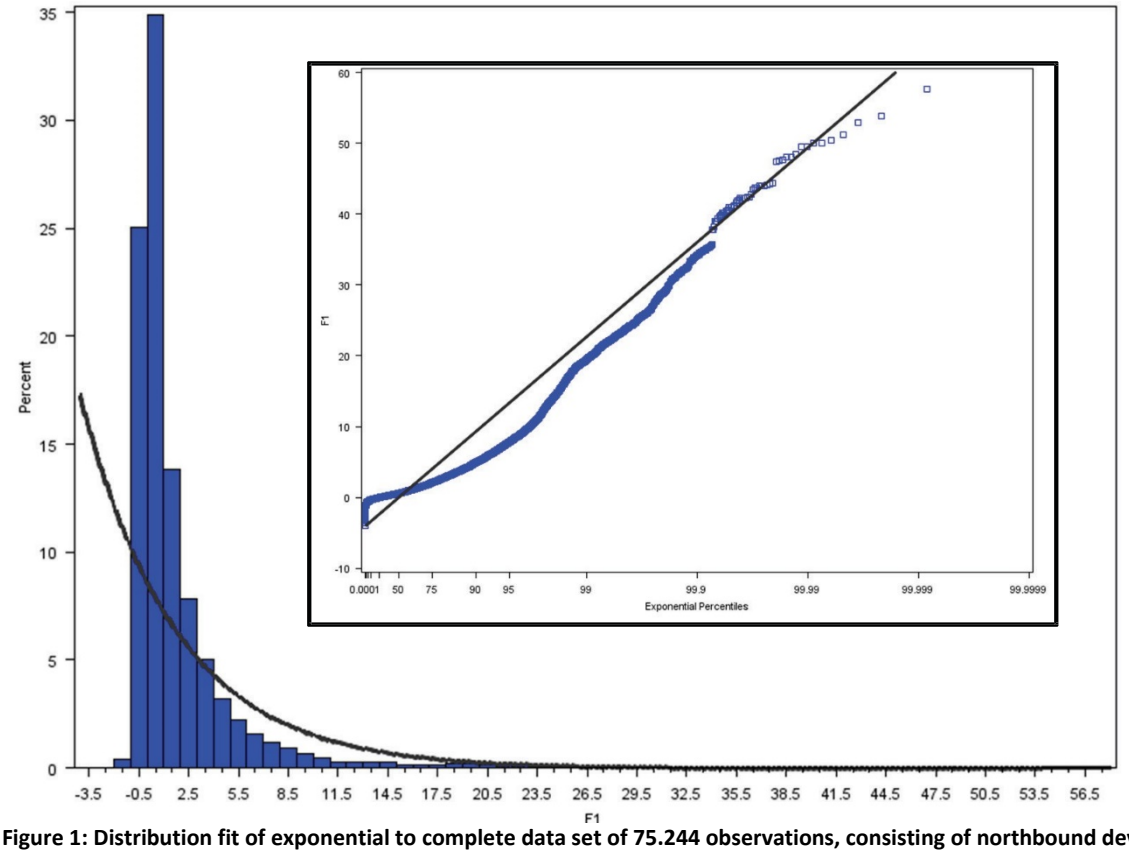

Figure 1: Distribution fit of exponential to complete data set of 75.244 observations, consisting of northbound deviations from timetable at eleven stations, weekdays, September to November 2014. Shown is histogram, fitted exponential distribution, and inset is probability plot. Horizontal axis "F1" is minutes deviation from timetable.

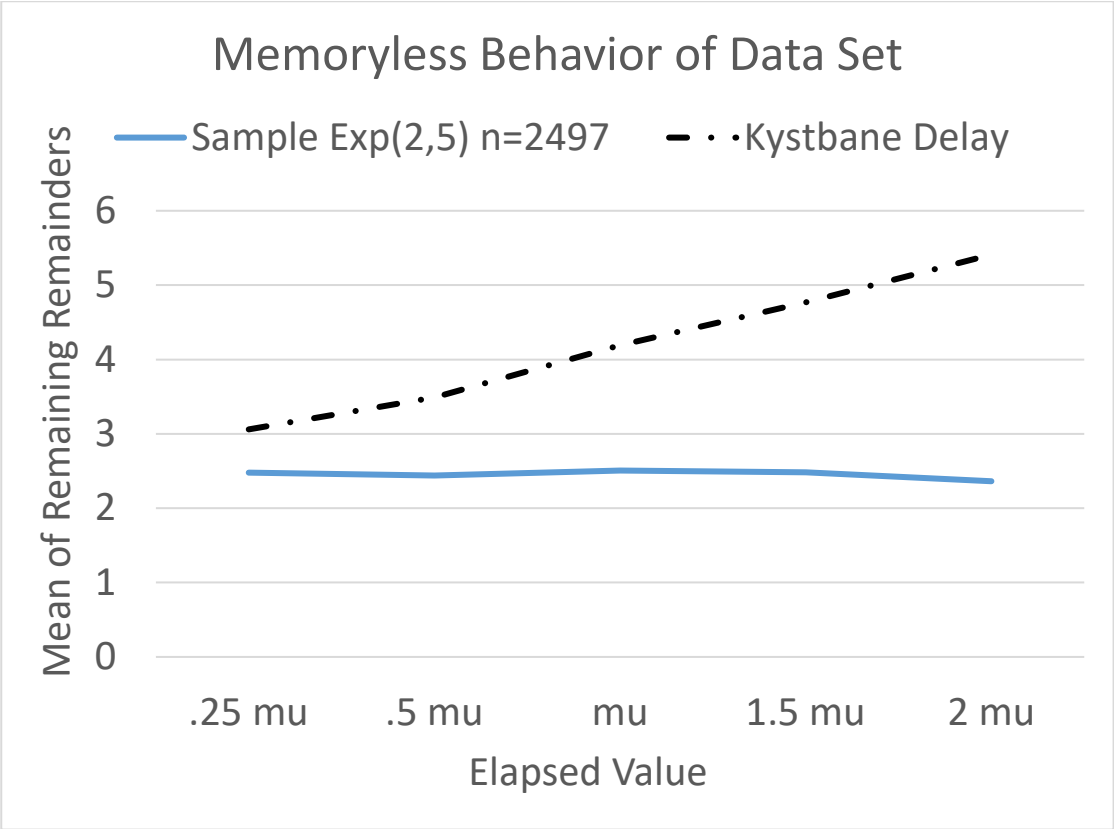

Figure 2: Test of memoryless property of Kystbane data. Vertical axis is mean value of remaining delay after value on horizontal axis scale has passed (which is marked in multiples of the whole data set mean). Solid line is a sample from an exponential distribution, and dashed line is actual result from sample data.

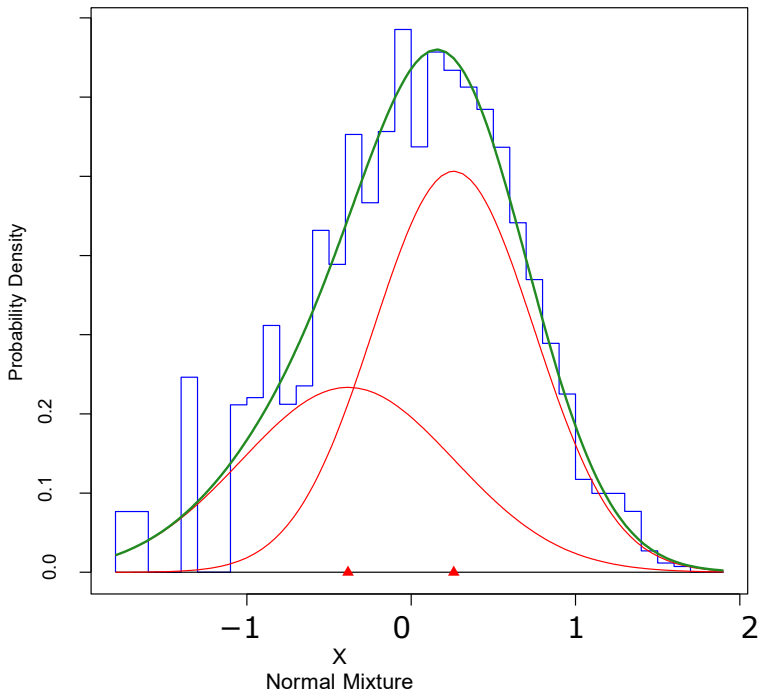

Figure 3: Mixed distribution fit to late departure data (sample size n=55.585, eleven stations on Kystbane). Data transformed to log base 10. Shown are histogram, summed distribution (green) and two component normal distributions (red). Triangle on x-axis is mean of component distribution.

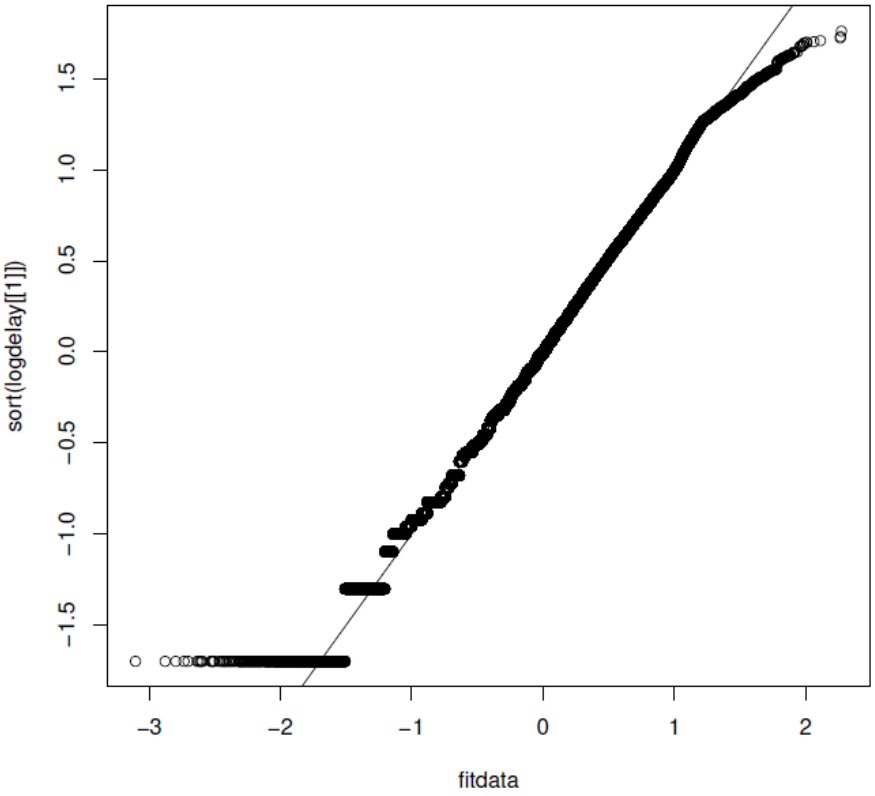

Figure 4: Probability plot of the late arrival sample data against the theoretical mixed distribution from Figure 3

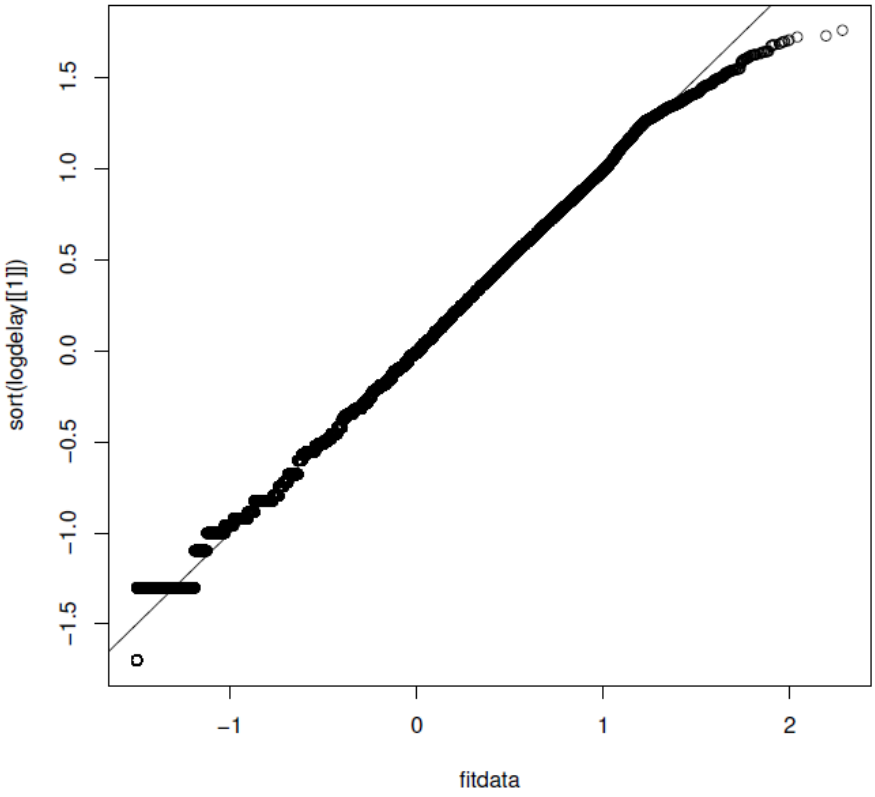

Figure 5: Probability plot of the late arrival sample data against the theoretical mixed distribution truncated to remove the lower tail consisting of on-time and early arrivals (x=-1,5). The point mass of the truncated data is not shown.

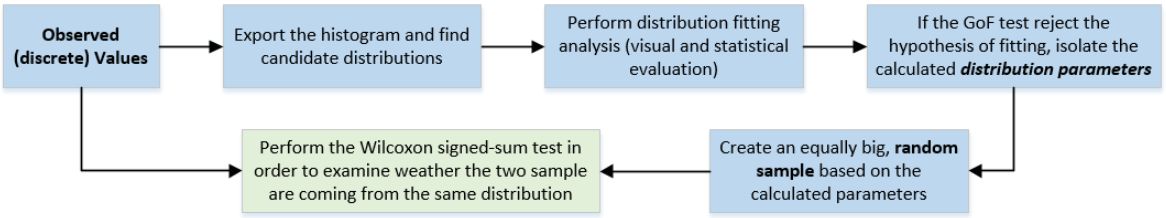

Figure 6: Methodology used to validate distribution fitting

References

Bolotin, V. A., 1994. Telephone circuit holding time distributions. In: Labetoulle, J., Roberts, J. (Eds.), Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks. Vol. 1. Proceeding of the 14th International Teletraffic Congress - Itc 14, Elsevier, pp. 125–134.

Browne, M. W., Cudeck, R., 1992. Alternative ways of assessing model fit. Sociological Methods and Research21 (2), 230–258.

Cerreto, F., Nielsen, B. F., Nielsen, O. A., Harrod, S. S., 2018. Application of data clustering to railway delay pattern recognition. Journal of Advanced Transportation, in press–.

Goverde, R., 2001. Statistical analysis of train traffic: The Eindhoven case. Tech. rep., The Netherlands TRAIL Research School, Delft.

Goverde, R., 2005. Punctuality of railway operations and timetable stability analysis, TRAIL thesis series no. t2005/10. Ph.D. thesis, Delft University of Technology, Delft.

Goverde, R. M., Hansen, I. A., November 2001. Delay propagation and process management at railway stations, document 175. In: Proceedings CD-Rom of the World Conference on Railway Research (WCRR 2001). Koln.

Hallowell, S. F., Harker, P. T., 1998. Predicting on-time performance in scheduled railroad operations: Methodology and application to train scheduling. Transportation Research Part A 32 (4), 279–295.

Harrod, S., Kanet, J. J., 2013. Applying work flow control in make-to-order job shops. International Journal of Production Economics 143, 620–626.

Johnson, R. A., Wichern, D. W., 2007. Applied multivariate statistical analysis (sixth edition). Pearson Education, Inc.

Jovanovic, P., Kecman, P., Bojovic, N., Mandic, D., 2017. Optimal allocation of buffer times to increase train schedule robustness. European Journal of Operational Research 256, 44–54.

Law, A. M., Kelton, W. D., 2000. Simulation Modeling and Analysis, 3rd Edition. McGraw-Hill, Boston.
Lindfeldt, A., 2015. Railway capacity analysis - methods for simulation and evaluation of timetables. Ph.D. thesis, KTH Royal Institute of Technology.

Massey, Frank J., J., 1951. The K olmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association 46 (253), 68–78.

Nie, L., Hansen, I. A., 2005. System analysis of train operations and track occupancy at railway stations. European Journal of Transport and Infrastructure Research 5 (1), 31–54.

Nielsen, O. A., May 2017. IPTOP - overview, integrated public transport optimisation and planning. In: Proceedings of Bane Konference (RailCPH), 2017. Banebranchen, Copenhagen. URL http://www.banekonference.dk/sites/default/files/slides/12/1500_IPTOP%20Otto%20Anker%20Nielsen%20v2.pdf

Richter, T., Landex, A., Andersen, J. L. E., November 2013. Precise and accurate train run data: Approximation of actual arrival and departure times. In: 10th World Congress on Railway Research. UIC, Sydney.

Vromans, M., 2005. Reliability of railway systems, TRAIL thesis series t2005/7. Ph.D. thesis, Erasmus University, Rotterdam.

Welch, N., Gussow, J., 1986. Expansion of Canadian National Railway's line capacity. Interfaces 16 (1), 51–64.

Wen, C., Li, Z., Lessan, J., Fu, L., Huang, P., Jiang, C., 2017. Statistical investigation on train primary delay based on real records: evidence from Wuhun-Guangzhou HSR. International Journal of Rail Transportation 5 (3), 170–189.

White, T., 2005. Alternatives for railroad traffic simulation analysis. Transportation Research Record 1916, 34–41.

Yang, Y., Li, J., Wen, C., Peng, Q., Lessan, J., 2017. Statistical distribution analysis of high-speed railway delay causes: Evidence from Guangzhou Railway Corporation in China. In: RailLille 2017 - 7th Interna- tional Conference on Railway Operations Modeling and Analysis. International Association of Railway Operations Researchers, Lille, France, pp. 1511–1531.

Yuan, J., Hansen, I., 2002. Punctuality of train traffic in Dutch railway stations. 10.1061/40630(255)73., 522–529.