M1 Project: Railway Delay Statistics and Modelling

Lusine Shirvanyan
Erasmus Mundus Master in Complex Systems Science,
Centre for Complexity Science, University of Warwick

We consider the problem of extracting statistical information regarding the distribution of train delays in the UK. Our work seeks to extend the work of Briggs and Beck [1], who showed that the q-exponential distribution accurately fits the distribution of observed delays for UK rail network. We studied the statistics of consecutive delays to understand the spatial and temporal correlation between the delays. The intended application of this research is to implement an accurate journey planning algorithm that can incorporate real-time data.

1 Introduction

The rail network in the UK consists of over 15,000 km of track. Planning journeys with a desired arrival time is a challenging problem, as not only are there often multiple routes between two points, to accurately predict journey time an understanding is needed of the reliability of the different parts of the network. Moreover, the likelihood of a delay can vary with the time, day of the week, as well as factors such as weather and line obstructions that can affect large parts of the network, creating a complicated dependency structure.

To help people plan their journeys, there are a number of route planning algorithms made available online. However, these are mostly limited—they plan journey under the assumption that no train is ever delayed! This is particularly problematic when a journey requires a change of trains, as even a short delay can result in a long delay in the event of a missed connection.

Although, statistics regarding the reliability of the different train routes are routinely gathered, they are normally too simple to be very helpful for route planning. For example, if a train route has 20 trains per day, and is 95% reliable, this mean that one train is cancelled, or it could mean that once every five days, five trains in a row are cancelled. The reliability is the same, but it makes a great difference to the distribution of journey times experienced by passengers.

This is why it is important to have an understanding of spatial and temporal correlations between delays. For example, if we are travelling from York to King Cross and the previous

train was delayed by three minutes, what is the reliability of trains departing over the next two hours?

Because of engineering works last for few weeks, it is desirable to be able to observe changes in reliability in the network over quite short time scales. It is also useful in the case of changes in timetabling to be able to adapt in a short period of time. Even though there is potentialy many years of data is available, it important to able to extract information from short amount of data. It is useful to be able to fit parametric models of data.

In their work, Keith and Beck showed that collected delays from 23 major stations in UK for the period September 2005-October 2006 can be accurately modelled by a two-parameter q-exponential distribution [1]. Their preliminary investigation implied that the following model (1) would fit well.

$$e_{q,b,c}(t) = c(1 + b(q-1)t)^{\frac{1}{1-q}} \tag{1}$$

Here t is the delay, $0 <$ q < 2 and b >0 are shape parameters, and c is a normalization parameter. Levenberg-Marquardt method of solving nonlinear least squares problems has been used for finding the best fit parameters for the distribution of delays. Figure 1 illustrates some of the results from Keiths and Becks work [1] for fitting model for all delays.

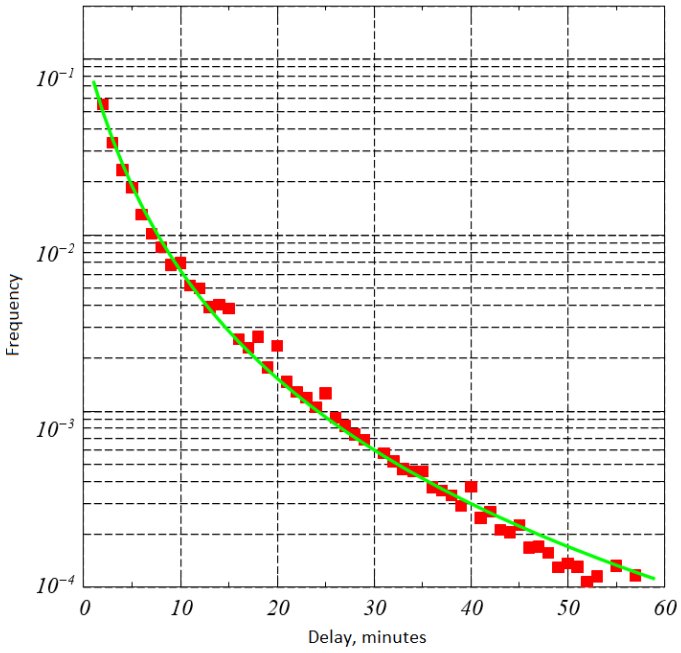

Figure 1: All train data and best-fit q-exponential: $q = 1.355 \pm 8.8 \times 10^{-5}$, $b = 0.524 \pm 2.5 \times 10^{-8}$ (by Keith and Beck[1])

The goal of this work is to extend understanding of the statistics of multiple trains. First of all we checked that new data for the period of time from 2014 to 2015 still fits weel the

q-exponential distribution. Then we focus on new types of data analysis, such as conditional probabilities and correlations between delays, because they can answer when a train is delayed, what is the probability that it is still delayed later in its journey? Or what is the probability that later trains on the same line are delayed? This will allow accurate simulation of passenger movement through a rail system with delays, the prediction of arrival times and maximization of departure time in a trip planning algorithm.

2 Data

2.1 The structure of the collected data

The advent of real-time train information available on the internet for the British network (http://www.nationalrail.co.uk/ ldb/livedepartures.asp) has made it possible to gather a huge amount of data .We collected data on departure times for 10 major stations for the period January 2014 to May 2015 using software which downloads the real-time information from webpage every minute for each station. Data are collected in YAML(Yet Another Markup Language) files and each file contains information about the departures from one particular station to the other stations for one day. Each row represents a list of pairs of departure time and the corresponding delay for the final destination. As each train eventually departs, the most recent delay value is saved to a database.

Figure 2 presents an example of such yaml file. According to it, the train from YRK to KGX departing at 07:01 from YRK has been delayed for 3 minutes, while the next one at 07:37 has been delayed for 5 minutes and etc.

```
# Last updated 2015-02-11 01:01
# Trains from YRK

ABD: {'0737': 3, '1154': 0, '1532': 0, '1555': 0, '1754': 0}
BHM: {'1936': 0, '2035': 0, '2045': 0}
BPN: {0719: 0, 0827: 0, 0918: 0, '1027': 3, '1118': 0, '1227': 5, '1318': 0, '1427': 0, '1518': 0, '1627': 0, '1718': 0, '1827': 0, '1918': 0}
BRI: {'1845': 0, '1945': 0}
BUY: {'0755': 0, 0845: 0, 0911: 0, '1011': 0, '1111': 0, '1211': 0, '1311': 0, '1411': 0, '1511': 0, '1611': 0, '1704': 0, '1811': 0, '1913': 0, '2011':
DEE: {'1732': 0}
EDB: {0855: 0, 0934: 0, 0937: 2, 0953: 4, '1054': 1, '1132': 0, '1136': 0, '1253': 0, '1332': 0, '1336': 3, '1454': 0, '1654': 0, '1836': 9, '1854': 0,
GLC: {0829: 0, '1032': 0, '1232': 0, '1432': 0, '1632': 0, '1736': 2, '1833': 9}
HUL: {'0730': 2, 0843: 0, '1019': 0, '1145': 0, '1247': 0, '1344': 0, '1502': 5, '1611': 0, '1725': 0, '1840': 2, '2229': 0}
INV: {'1355': 0}
KGX: {'0701': 3, '0737': 5, 0802: 0, 0820: 0, 0831: 0, 0856: 2, 0931: 28, 0957: 10, '1003': 6, '1027': 0, '1031': 5, '1059': 0, '1131': 0, '1157': 6, '1
LDS: {'0740': 8, '1729': 0, '2020': 3, '2123': 0, '2148': 0, '2213': 0, '2313': 0}
LIV: {'0715': 2, '0737': 3, 0815: 0, 0840: 0, 0915: 5, 0941: 0, '1016': 3, '1041': 7, '1115': 0, '1141': 0, '1216': 12, '1241': 0, '1315': 0, '1341': 0,
LMS: {'0727': 7, 0826: 0, 0935: 3, '1035': 6, '1135': 0, '1235': 3, '1335': 0, '1435': 0, '1535': 0, '1606': 0, '1735': 0, '1835': 0}
MAN: {'1553': 7, '1753': 0, '1853': 5, '2230': 0}
MBR: {0718: 0, 0815: 0, 0915: 0, '1015': 0, '1115': 0, '1215': 0, '1315': 0, '1415': 0, '1515': 0, '1615': 0, '1715': 0, '1815': 9, '1916': 2, '2016': 3
MIA: {'0652': 0, '0724': 0, '0750': 2, 0824: 0, 0853: 0, 0924: 2, 0953: 6, '1024': 4, '1053': 2, '1124': 0, '1153': 0, '1224': 0, '1253': 0, '1324': 0,
NCL: {0708: 2, '0732': 0, 0808: 0, 0835: 0, 0850: 0, 0908: 0, 0948: 3, '1008': 5, '1035': 0, '1048': 0, '1108': 0, '1148': 0, '1208': 0, '1235': 0, '124
PLY: {'0743': 2, 0845: 0, 0945: 6, '1045': 0, '1145': 0, '1545': 0, '1645': 0, '1745': 0}
PNZ: {'1245': 0, '1345': 0, '1445': 0}
SBY: {'1102': 0, '2013': 0}
SCA: {'0740': 1, 0840: 7, 0940: 2, '1040': 0, '1140': 0, '1240': 2, '1340': 3, '1440': 0, '1540': 4, '1640': 0, '1740': 6, '1840': 6, '1940': 0, '2040':
SHF: {'1106': 0, '1502': 2}
SUN: {0958: 3, '1322': 2, '1451': 0, '1841': 7, '2119': 0}
Whitby Bus Station: {0821: 0, 0927: 0, '1047': 0, '1307': 0, '1447': 0}
```

Figure 2: Departures from YRK to 25 other stations on 2015-02-11 collected in a yaml file.

2.2 Limitations of the collected data

There were some limitations of the representation of the collected data, which turned out to be an obstacle for some of the analysis. First of all, the fact that data are spread between many files makes it difficult to collect all the departure data for a particular station for a specific period of time. In order to obtain this data, we would have to scan all the files. This would be time consuming since only 10% or even less of data is relevant to a given station as the final destination. Another problem was the format of departure times in the file, which, as we can see in Figure 2, is not the standard format for date/time. This was not allowing us to make queries for collecting departures for some particular times of the day, for example, looking at delays in off-peak times only or comparing delays of consecutive trains. We managed to overcome these limitations by transferring all data from yaml files to SQLite database and keeping all data in an easy to use format. Before giving more details about the SQLite generated database, we should focus our attention to some limitations of data.

As mentioned before, each file contains information about departures from one source station to the final destinations only and there is no information about intermediate stations for each journey. This lack of information did not allow us to analyze the correlations between delays in different stations from the same journey. This resulted in setting the additional task of train identification, which will be discussed in more detail in section 3.1. And the last limitation is that only departure times are provided in the data, with no information about arrival times, which had some impact on the train identification task.

2.3 Generated SQLite database

Because of the limitations of data representation mentioned in subsection 2.2 and its implementation advantages, SQLite database was chosen to store the data. As we can see in Figure 3, three tables were constructed for storing the data from the yaml files. Stations table contains names of all stations collected from all files and additionally assigns unique StationID to each of them, which makes querying of data much faster, by avoiding string comparisons every time. The Departures table keeps all departure records from the files, by assigning unique DepID to each of them. Corresponding ID-s from the Stations table are recorded for the source and the destination stations (marked in green in figure). Additional Weekday column was added to table, which contains information about the day of the week extracted from the date, which makes possible the analysis considering the day of the week. The Trains table was constructed to store the results of the train identification task.TrainID and JourneyID in the table are used for grouping different records from the Departures table into the same journey or same type of train.

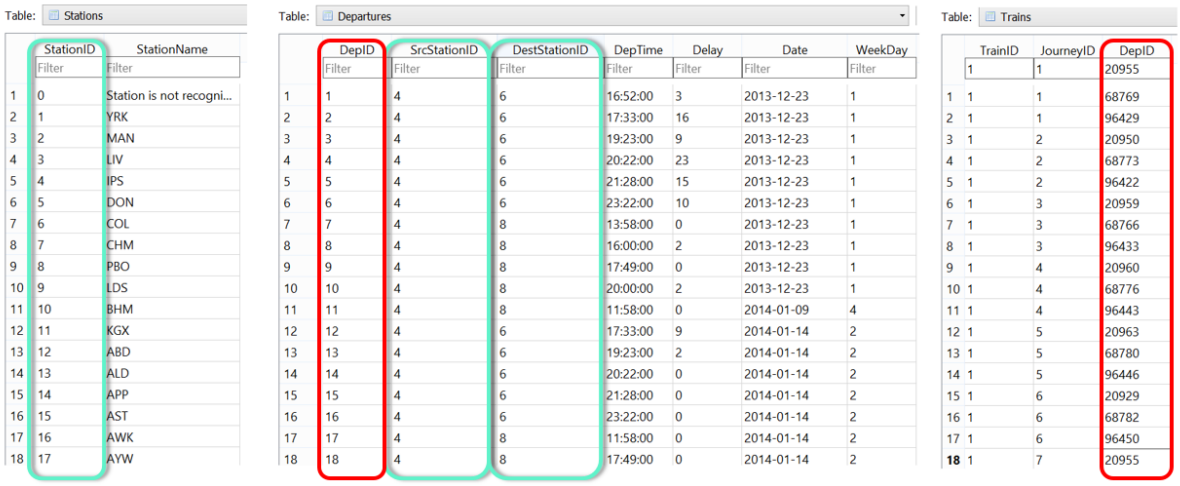

Table: Stations

	StationID	StationName
	Filter	Filter
1	0	Station is not recogni...
2	1	YRK
3	2	MAN
4	3	LIV
5	4	IPS
6	5	DON
7	6	COL
8	7	CHM
9	8	PBO
10	9	LDS
11	10	BHM
12	11	KGX
13	12	ABD
14	13	ALD
15	14	APP
16	15	AST
17	16	AWK
18	17	AYW

Table: Departures

	DepID	SrcStationID	DestStationID	DepTime	Delay	Date	WeekDay
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	4	6	16:52:00	3	2013-12-23	1
2	2	4	6	17:33:00	16	2013-12-23	1
3	3	4	6	19:23:00	9	2013-12-23	1
4	4	4	6	20:22:00	23	2013-12-23	1
5	5	4	6	21:28:00	15	2013-12-23	1
6	6	4	6	23:22:00	10	2013-12-23	1
7	7	4	8	13:58:00	0	2013-12-23	1
8	8	4	8	16:00:00	2	2013-12-23	1
9	9	4	8	17:49:00	0	2013-12-23	1
10	10	4	8	20:00:00	2	2013-12-23	1
11	11	4	8	11:58:00	0	2014-01-09	4
12	12	4	6	17:33:00	9	2014-01-14	2
13	13	4	6	19:23:00	2	2014-01-14	2
14	14	4	6	20:22:00	0	2014-01-14	2
15	15	4	6	21:28:00	0	2014-01-14	2
16	16	4	6	23:22:00	0	2014-01-14	2
17	17	4	8	11:58:00	0	2014-01-14	2
18	18	4	8	17:49:00	0	2014-01-14	2

Table: Trains

	TrainID	JourneyID	DepID
	1	1	20955
1	1	1	68769
2	1	1	96429
3	1	2	20950
4	1	2	68773
5	1	2	96422
6	1	3	20959
7	1	3	68766
8	1	3	96433
9	1	4	20960
10	1	4	68776
11	1	4	96443
12	1	5	20963
13	1	5	68780
14	1	5	96446
15	1	6	20929
16	1	6	68782
17	1	6	96450
18	1	7	20955

Figure 3: Database generated from Yaml files.

3 Results

3.1 Train identification

We grouped departures for the same final destination in such order, that with high probability its the same train that goes through all that stations to the same destination. Obviously, this could not be done using only the data we have, since there is no information such as distance between stations and the route trains can take to the final station. Therefore we gathered information about the different types of trains for one particular journey and we then tried to match these to the departures from the database, by providing some basic properties about each type of train, such as in which stations the train stops and what the difference between departure times for two consecutive stations. Figure 3 illustrates the result of train identification for journeys from YRK station to KGX station. We identified 3 different types of trains going from YRK to KGX, 2 of which have stops in DON and PBO stations (red and black), whereas third type(blue) stops in PBO only. On the left side is a plot for data collected manually from the website for 01-06-2015, and on the right side are plotted trains identified from data set for some 07-05-2015. Most of the trains were identified correctly, however, there were some misidentified trains too. In addition, we can notice that we do not have last, final station in our second graph, which is because there is no information provided about arrival times. However, the results were good enough for using them for correlation analyses.

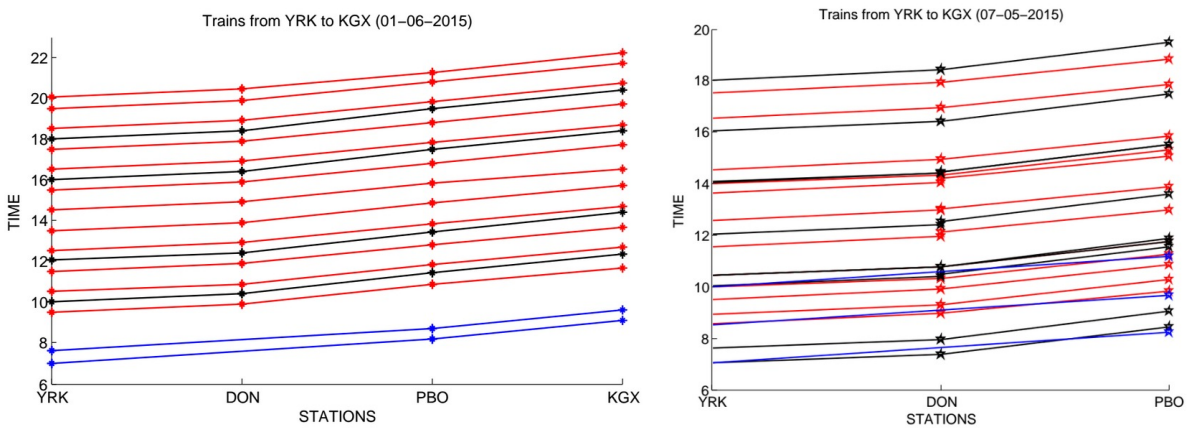

Figure 4: Train identification. On the left different trains and journeys collected manually. On the right side trains/journeys identified from our data.

3.2 Correlation between delays in different stations

Now, when we have some information about departures from same journey, we can look at correlation of delays between consecutive stations. Scatter plot in Figure 5 illustrates correlation of delays for each pair of stations from YRK to KGX. As we can see, there is mainly a positive correlation between all stations, which is quite natural, since when a train is delayed in one of the stations, it is with a high probability that it was either delayed previously on its journey, or it will be delayed later. However, we see some big delays on the plot, which did not lead to delay in next station, which is very unlikely in reality. This probably means that those delays are from misidentified trains, i.e. departures grouped together as one part of one journey are independent in real.

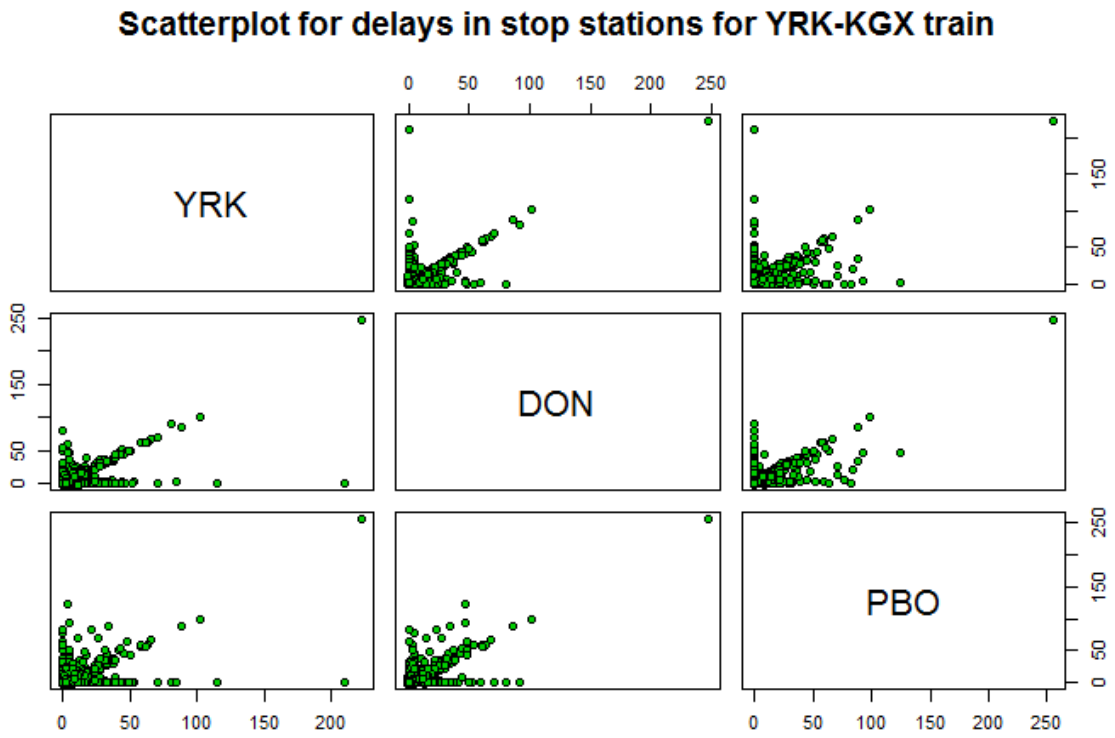

Figure 5: Correlation between delays in different stations of same journey.

3.3 Distribution of delays for the days of the week

Another thing we investigated is the distribution of train delays for the different days of the week. The violin plot in Figure 6 presents clearly the differences and similarities between delays for each day of the week. For all days the mean delay is approximately 4 minutes, whereas it's slightly more (about 2 minutes) for Monday. Another noteworthy fact is that the probability of big delays on Sunday is much higher comparing to all other days, while there is much smaller range of probable delays on Wednesday. Overall, the range of delays decreases at the beginning of the week and and then increases after Wednesday. The reasons of this phenomenon can be various, however, the important thing here is that taking into account the day of the week may help doing more accurate predictions about delays.

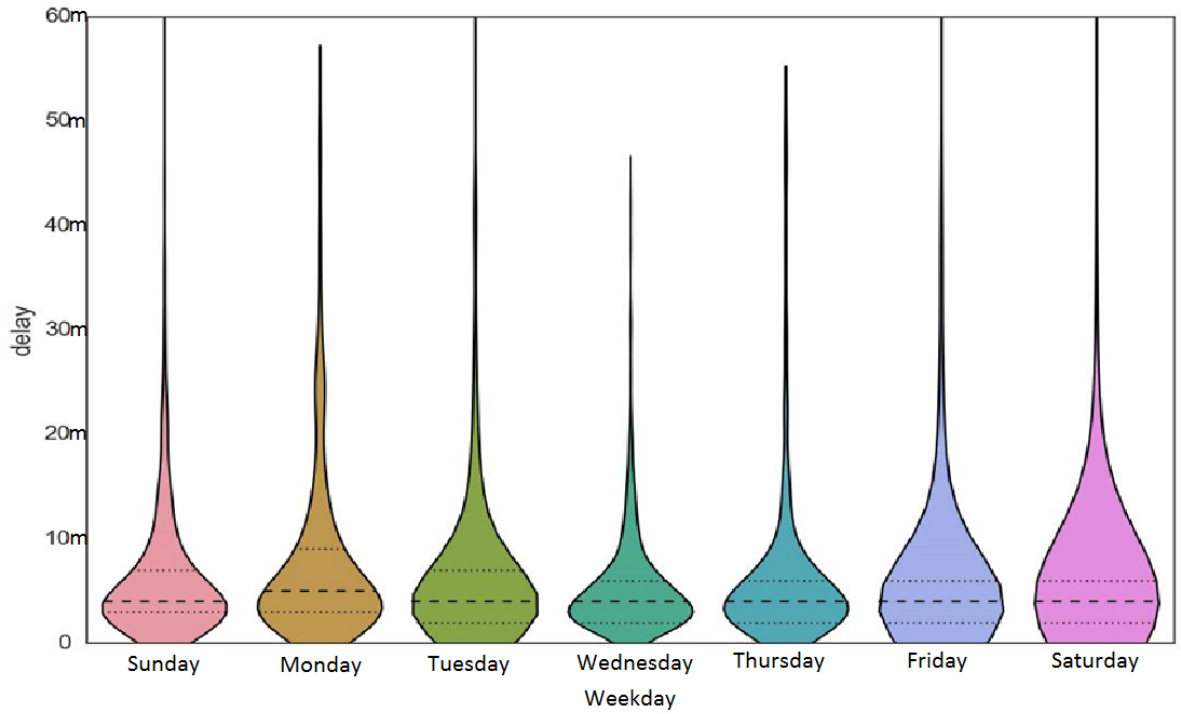

Figure 6: Distribution of delays for the days of the week.

3.4 Conditional probabilities

The last thing we analyzed in the project was the conditional probabilities of delays. More specifically, we looked at the probability distribution of the delay of next train conditional on the previous train being late by n minutes.

Because of small amount of data this distributions are fairly noisy, and it is difficult to visualise using a standard plot. To show how the distribution is changing with n, we have plotted the differences between distributions and the $n = 0$ distribution, see Figure 7. The colour, from blue to red, indicates increasing n. You can see that as n increases, the probability of having no delay for the subsequent train decreases, with a corresponding increase in the probability of having a positive delay. Here we are focusing on the probabilty of small delays, to understand longer delays, one could model the conditional distribtions as q-exponential distributions.

We tried using singular value decomposition (SVD) for smoothing the distributions. Formally, the singular value decomposition of an $m \times n$ matrix M is a factorization of the form $M = U\Sigma V^*$, where U is an $m \times m$, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and V^* (the conjugate transpose of V, or simply the transpose of V if V is real) is an $n \times n$ real or complex unitary matrix. We get rid of noise which is high rank, by just leaving meaningful information. The method of validation have been applied for minimization of negative log likelihood, by using 80% of data for training, and

the rest for testing. The plot in Figure 8 demonstrates that the rank of 7 minimises negative log likelihood the most, thus this value is the best for smoothing data. We also tried to smooth data by splitting it to different times of the day (e.g. morning, afternoon, evening), but it did not improve the result that we had.

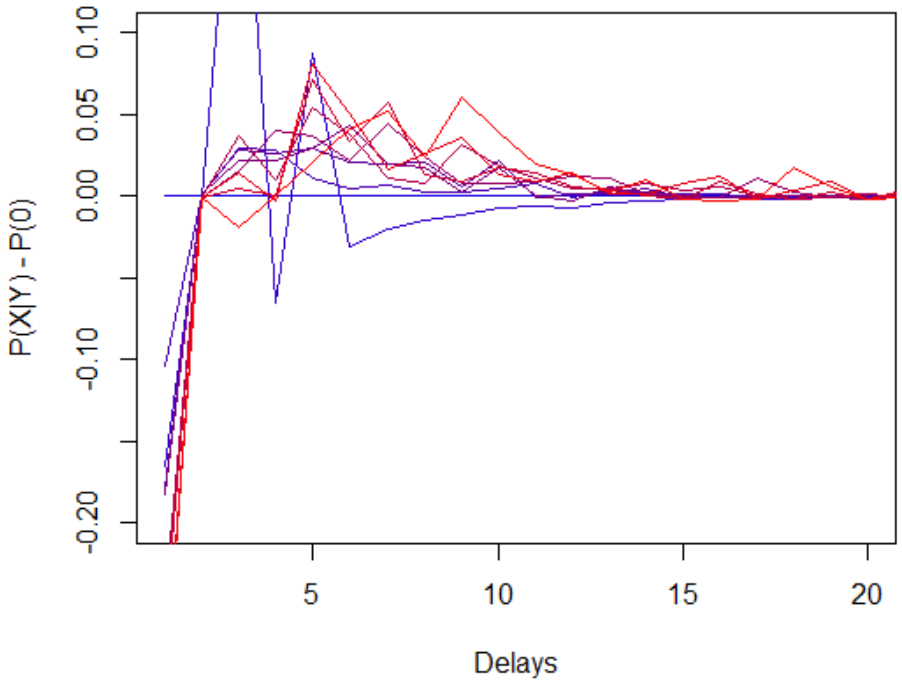

Figure 7: Conditional probabilities of delays for consecutive trains from IPS to COL. Subtracting the zero delays distributions normalizes the conditional probabilities by allowing us to compare different conditional distributions.

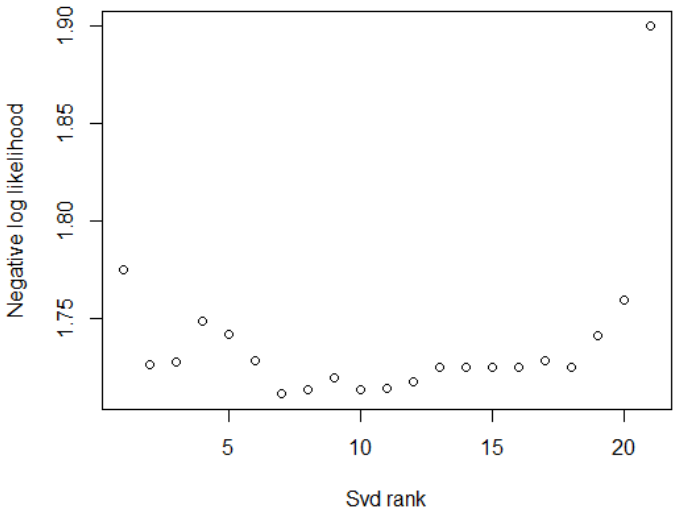

Figure 8: Negative log likelihoods after smoothing data using different ranks in SVD decomposition

4 Conclusions

We have presented analysis of train delays distribution in the UK rails system. First of all, although we were able to identify the majority of the trains for a particular journey by providing basic properties of the journey, there were misidentified trains. Because of that, the data analysis can not be very accurate. Possible solution to this could be providing more details about each journey as an input, which would leave implications on the computations. Alternatively, the possibility of collecting more information about each departure (such as arrival times and stop stations) should be considered. Concerning the other statistics of the data set, we were able to find high correlation (0.76) between delays of consecutive stations during one journey despite the presence of misidentified trains. In other words, if a train is delayed, then with high probability it will be still delayed later in its journey. Also, it was found that distribution of delays is different for the different days of week, which means that it may be reasonable considering the day of the week in journey planning. These results may allow more accurate simulation of passenger movement through a rail system with delays, which was one of the main purposes for this project.

Acknowledgement

My gratitude goes to my supervisors Dr. Keith Briggs from BT and Dr. Ben Graham from Warwick University for their guidance and help provided during the project.

Reference

1. Keith Briggs and Christian Beck. Modelling train delays with q-exponential functions (2007)

2. Keith Briggs and Peter Kin Po Tam. Optimal trip planning in timetabled transport systems possessing random delays(2011)

3. Nonlinear-Least-Squares Analysis of Slow-Motion EPR Spectra in One and Two Dimensions Using a Modified LevenbergMarquardt Algorithm

4. Shmuel Friedland University of Illinois at Chicago. The Role of Singular Value Decomposition in Data Analysis

5. Kirk Baker March 29, 2005 (Revised January 14,2013).Singular Value Decomposition Tutorial

6. J.-F. Bercher and C. Vignat.A new look at q-exponential distributions via excess statistics