Swiss railway delay analysis

Maria Pandele
maria.pandele@epfl.ch

Dimitri Lallement
dimitri.lallement@epfl.ch

Ejub Talovic
ejub.talovic@epfl.ch

Santiago Anton Moreno
santiago.antonmoreno@epfl.ch

Abstract

Transport is a key factor in today's society. We have to move often and fast in our daily lives. We also need to consider the impact some types of transport have on our ecosystem.
Public transport and the train system in Switzerland is constantly trying to improve for us to be able to move efficiently. The main issue that arise in every train system is the schedule and the delay the passengers have to cope with. It's one of the drawbacks that deter some people from using trains instead of their cars.
Our goal is to investigate where are the problematic stations located and quantify the delay depending on the location and time of a train but also to provide a tool that allows the user to predict a delay for his ride.

1 Introduction

The SBB division of passenger transport offers open data about train arrivals, departures, delays, etc.We want to use this data to understand the network of railway stations in Switzerland.

The main task was to analyze delay patterns. After studying some basic statistics like the proportion of departure with delays, we decided to also analyze what causes the delays. Another important part of the project was to make a model that predicts the probability of delay based on the hour of the day and the station you are in.

On top of all, we also made network graphs that represent Swiss trail network.

2 Data collection

First we used the open data sets of the CFF company. You can find it in this link.

However, this site only provides data for the preceding day. So we contacted the SBB CFF company directly and they provided us a full data set, with data starting around the beginning of 2018 up to today.

We decided to use the data set we got from the open data of the CFF company for our analysis. This was much more convenient for us because we started using the open data first and the full data set is too big to fit the memory of our laptops.

3 Datasets description

Our first and most important data set was "ist-daten-sbb.csv". Each row of this data set contain information about one ride, the most relevant being the departure and arrival stations, geolocation, planned departure and arrival time and the real departure and arrival time. It has 61716 rows. The column names were in German so we had to translate them in English. We also removed some columns that were irrelevant for us. We used this data set to generate a new data frame, Whose rows represent a line in the network. Meaning that it contains a list of station and the list of departure/arrival times for each ride.

We also used the data set "rail-traffic-information.csv". This data set is in English. It contains the most important information on rail traffic in Switzerland according to CFF. This data set has 14377 rows. A lot of data preprocessing has been done on this data set so you can read about it on the Reasons for the delays subsection.

The dataset we got from SBB contains data from January 2018 to September 2019. It has almost the same format as the one we used from SBB open data website except for few columns. It is quite a huge dataset having around 145GB because it also contains arrival/departure information for regional transportation networks and foreign railway companies. So our first step is to disregard these rows the dataset. Next, we noticed that

some columns are missing: ankunftsverspatung (arrival was delayed), abfahrtsverspatung (departure was delayed), geopos (longitude and latitude of the station). The first two of them are trivial to add and compute. For the geoposition one, we loaded another dataset which has information about the transportation stops in Switzerland from this link (around 37980 rows) and extracted the ones we were interested in. In some cases the names were different: stations contain the abbreviation of the canton (Wil SG), had special characters (Ambrì-Piotta), or were incomplete (Puidoux and Puidoux-Chexbres point to the same station). Luckily, we found 24 of these corner cases and we manually corrected them.

4 Delay analysis

4.1 Overview

For both departure and arrival delay we computing by subtracting the *planned departure time* to the *actual departure time*. Then we remove the negative values because they represent the trains that left early.
We chose to only analyze the departure delay because it is the one that interests users the most. After plotting the first distribution with a log we observe that most of the delayed trains are less than 3 minutes late. We chose to redefine the limit of the delay since this type of delay is not too penalizing for the users.

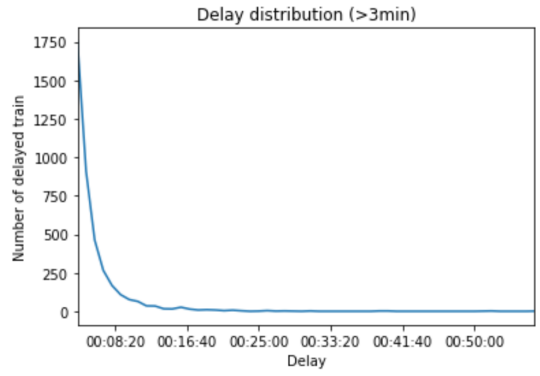

Figure 1: Delay distribution for train which are more than 3 min late

However even after this filtering, we see that the distribution of delays follows a power law distribution (Figure 2). The proportion of delayed train that are more than 3 minutes late is only 6.49 % of the total delayed trains.
We also checked if a late departing train necessarily have a arrival delay by computing the correlation between the two delays (0.57, less than what we expected). That means even if the train is delayed at the beginning it's not necessarily late to the arrival (it can accelerate, etc.).

4.2 Delay analysis through the network

We also wanted to check if the delays were uniformly distributed throughout the network or if some regions were more affected than others. To do that we extracted the longitude and latitude from the geolocation column.
Then we calculated the maximum delay for each section connecting two stations and plot the result on a map generated with openstreetmap.

4.3 Delay analysis per station

We want to investigate which stations are most affected by the departure delays. For this, we define the delay frequency in one station to be the percentage of delayed trains in one day. We plot a histogram with these frequencies and we find that in more than half of the SBB stations the number of delayed trains is between 0% and 20%.

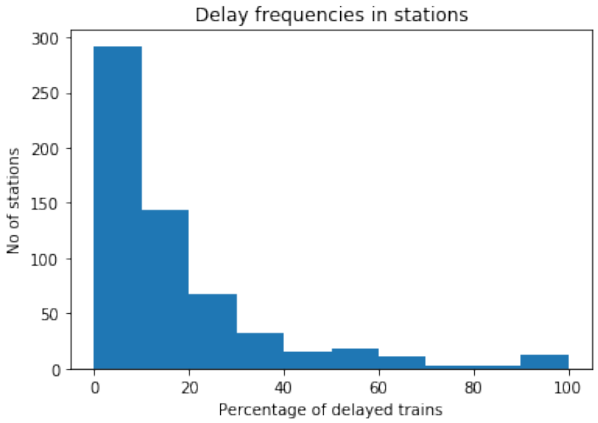

Figure 2: Histogram of delay frequencies in SBB stations

4.4 Reasons for the delays

As we are doing a data project on delays of the CFF trains, we tried to get some insight on what causes the delays. We did this using the dataset "rail-traffic-information.csv".

The column that were relevant to us are 'Title', 'Description' and 'Published', the last one giving us when the delay announcements was published. By looking carefully at the data we can see that a lot of delay information appear multiple times

within minutes/hours and the description will tell that cancellations and delays are expected (not a precise number), which are factors that can bias our conclusions and make it hard to quantify exactly how much each delay reason impacts the traffic. Therefore even tough this data set also contains columns that show us information between or near which stations the delays/cancellations might happen, we did not use it to try to quantify delays and cancellations for each station as we can use our "ist-daten-sbb.csv" data set for this purpose. Nonetheless, this data set allowed us to see the big picture of what causes the delays.

Sometimes, a delay or cancellation was reported with no reason or cause specified. We checked that a reason was specified 85% of the time, which is enough to have a meaningful data representation.

5 Delay prediction tool

In this section we mainly used existing methods from scikit-learn library (Sci, 2011). We wanted to have models that could predict if a train is delayed. Given the size of the dataset we considered simple models since more data is often better than complex models. Firstly we had to preprocess our data to be usable in those models since most models can only take numerical features as input and we had lot of categorical data or datetime objects (e.g. station name, arrival time). Thus we decided to use dummy variables for the station names splitting that feature in 600 different features for every station. Then we split the planned_arrival_time feature in 3 feature consisting of the hour, minutes and day.We splitted our dataset in a test and train set with a 30/70 proportion. considering only those features we fitted a Linear model with Tykhonov regularization like in equation 1 to try to predict the minutes of delay of a given train.

$$\min_{\beta \in \mathbb{R}^n} \|\beta X - Y\|_2 + \alpha \|\beta\|_2 \qquad (1)$$

Given the size of our dataset and the number of features, regularization seamed unavoidable and indeed it was since we ran a 5-fold cross-validation with the basic mean squared error on the regularization parameter and it gave us $\alpha = 2$. Then we wanted a model that could tell us the probability of a train to be delayed. Instead of a regression model, there we needed a classification model. We choose to fit a logistic regression with regularization like in Equation 2 to our data with the same features as for the linear regression and with the binary labels for each departure that encodes if the train has more than three minutes of delay.

$$\min_{\beta \in \mathbb{R}^n} \left\| \frac{1}{1 + exp(-\beta X)} - Y \right\| + \alpha \|\beta\|_2 \qquad (2)$$

We then code a function that lets the user input a station and a time and outputs the probability of delay computed by our logistic model and the arleady implemented predict_proba method.

```
enter the start station: Renens
enter the time(format d HH:MM): 2 18:00
Probability of delay:0.13
```

Figure 3: Example of a request of prediction

6 Results

6.1 Delay analysis

6.1.1 Delay pattern

One our goal was to discover potential patterns in terms of delays. By studying the delay distribution histogram according to the time of day (Figure 4) we can observe some interesting results.

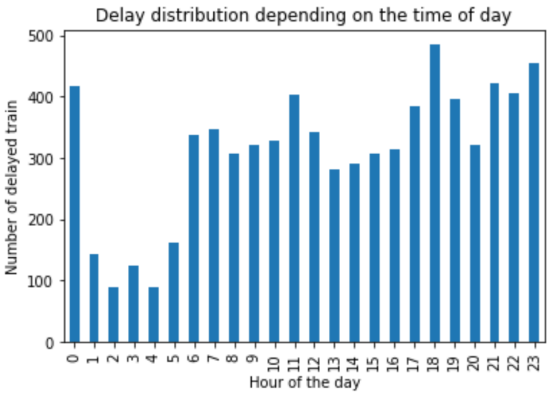

Figure 4: distribution of delays according to hours of the day.

Firstly, as expected the number of delays is lower at night (the train frequency is lower therefore the number of delays too).
However, what is more surprising is that we do not observe peaks during rush hours. There is an increase in delay around 6 p.m. but it is less significant than what we expected to see (almost as many late trains as at 11 p.m. for example).

6.1.2 Delay through the railway network

By plotting the maximum delay for each section of the network (Figure 5), we observe that there are no lines particularly affected by significant delays (more than 15 minutes).

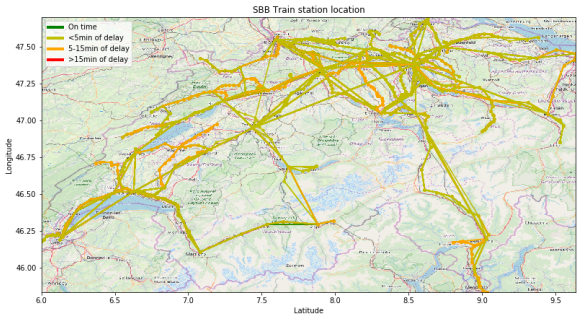

Figure 5: Overview of the network delay, each route is colored according to its maximum delay.

Some line trajectory seem quite strange but it's justify because the train is a direct one. We only have the departure and arrival station so we can't represent the real trajectory.

6.1.3 Delay per station

We also wanted to investigate in which stations the delays are the most frequent. We observe that few stations are affected by delays: Mosen(100%), Corgémont(92.30%), Grandval(80.55%) etc. None of the large cities in Switzerland have a significant number of delayed trains: Bern (7.81%), Lausanne (6.06%), Geneva (8.5%), Zurich (3.87%).

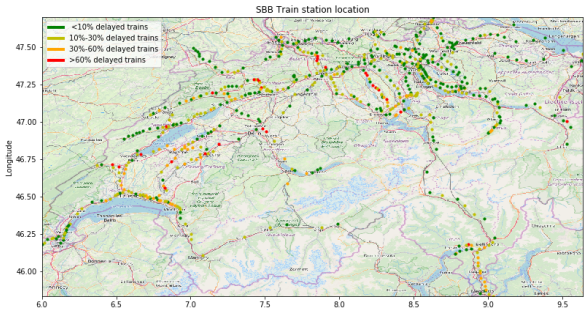

Figure 6: Distribution of delays according to the departure station.

6.2 Reasons for the delays

In Figure 7, we can see the proportion of the time a delay cause was reported. We can observe that most delays are caused by technical faults, construction work and overhead line problems, which

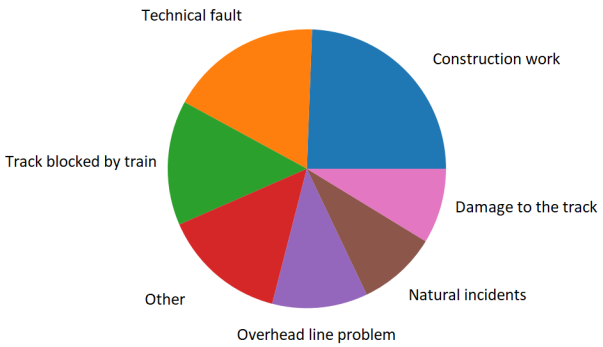

Figure 7: Delay causes

are all somewhat recurrent technical problems that have room for improvement.

6.3 Delay prediction tool

The Linear model we fitted with the optimal alpha equal to 2 gets a test mean squared error of 2.7 and a test mean absolute error of 0.9. This means that in average we are 1 minute off from the real delay which seems like a really good score considering the simplicity of our model.
In Figure 8 we see scores for the logistic model. The accuracy of our model is great but we should be careful in deducing that our model is great because the two classes are heavily skewed. Indeed there is only 12% of positive cases so if it basically gives every train a negative classification it will have a 0.88 accuracy like we already have. But our model is not a naive one has we see it at tries to predict delayed trains. As we see the recall for the $\boldsymbol{True}$ class is not that great, that can be explained by the skewedness of our data. We tried to fit the logistic with unbalanced weights but it did not give significant improvement

	precision	recall	f1-score	support
False	0.890	0.989	0.937	14816
True	0.659	0.145	0.238	2128
accuracy			0.883	16944
macro avg	0.774	0.567	0.587	16944
weighted avg	0.861	0.883	0.849	16944

Figure 8: Score for the Logistic Regression

7 Conclusion

This project allowed us to have a fairly global vision of delays in the Swiss railway network. We

answered all of the research questions we listed in the previous milestone. The future of the project would be to apply the same analyzes on the complete dataset provided by SBB.

References

[Sci2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. 2011. *Scikit-learn: Machine Learning in Python*, *Journal of Machine Learning Research* 12:2825-2830.