CS698H: An Analysis of Delays of Trains

Arindam Sarkar (16111004), Subhadip Nandi (16111023)

***Abstract*—In this project, we analyse delays of trains of Indian Railways, and try to find out if junction traffic (average number of trains passing) and number of routes via junction provide an explanation for delays.**

I. Introduction

Train delays have come to be associated as a characteristic of Indian Railways. Common speculations are endless - ever-increasing number of trains, infrastructure, chain pulling and so on. As a part of the project we try to find out how (or if) train delays are related to junction/station characteristics like number of trains passing through the junction on an average (junction traffic) and number of routes via junction.

II. Data Gathering

For any statistical analysis of this sort, we need data to back our hypotheses. We obtain the data by scraping various websites (cleartrip.com, etrains.info, irctlive.com,wikipedia). For scraping we have used python libraries - BeautifulSoup, urllib, dryscrape.

The reason for scraping multiple sites was that, we needed different sorts of data - list of stations, list of trains, delay of trains, number of routes per station, and no site provided all the data.

A primary challenge was to collate the data together - because every site has a different format, only relief being the prevalent use of train-codes and station-codes what acted as a bridge in the task.

The process of collation also introduced some ”interesting” errors - for instance when we tried to collate wikipedia data for number of routes with station codes, we had to do an approximate search for station names, which resulted into matching first instance of a station name when there are several stations in a city!

But with some hacking around, we were able to gather a list of stations (4k+), trains passing through them (with detailed schedule, but of-course some missing data), and delay of each train for the last year.

III. Analysis

As part of analysis we asked the following questions:

- Is there a correlation between Traffic through a junction and avg. delay of trains passing through the junction?
- Is there a correlation between Number of Routes via a junction and avg. delay of trains passing through the junction?
- How good are Traffic and No. of routes as predictor for Delay?
- Does average delay over stations with high traffic/ larger no. of routes differ significantly from stations with low traffic/ smaller no. of routes?
- Is delay on Monday significantly different from delay on a weekend?
- Also, we compare two stations (specific ones, just for sake of experimentation) and see if difference in delays is significant.

Also, we tried out if excluding winter months when there is Fog in much of Northern India can affect results (roughly Dec-Feb).

IV. Findings

We found that correlation between Station Traffic and avg. delay of trains passing through junction is quite low (0.0349). For No. of routes passing and avg. delay of trains, correlation was 0.177, which is higher than correlation with traffic, but pretty low.

Then we perform linear regression with traffic and no. of routes via a junction/station as predictor variable (variables scaled properly), and delay as prediction. We get pretty low $R^2 = 0.037$ indicating that there is a lot of unexplained variance in model, that is our predictor variables are not really great at explaining delay!

But F_Stat = 4.211, and P(F_Stat) = 0.016, so we can reject the null hypothesis that the fit of the intercept-only model and obtained model are equal (5% significance level). Again, looking at the coefficients and comparing p-values, we observe that we can’t reject null hypothesis of coeff. of avg. traffic being 0, while we are able to reject null hypothesis of coeff. of no. of routes being 0.

So, although, model isn’t a great fit for the data, no. of routes seem to explain delay more than traffic.

We also performed a 2-sample test by taking 60 random high traffic junctions and 60 random low traffic junctions, and compared avg. delays across them, but we didn’t observe any significant result, and were unable to reject the null hypothesis that avg. delays are same. We repeated the same experiment for high/low no. of routes, and ended up with similar results.

We also checked day-wise (avg. of all trains across all 7 days), and tried 2-sample test for pairs (Monday-Friday, Monday-Saturday, Monday-Sunday), but we couldn’t conclude that average delays are significantly different.

As another experiment, we took two stations in particular - Bathinda and CoochBehar (just out of thin air!), and performed a 2-sample test, and were able to reject the null hypothesis that avg. delays across them are same.

We also tried out a few of above for months excluding winter months, but results differed only marginally, specifically, delays were more on an average.

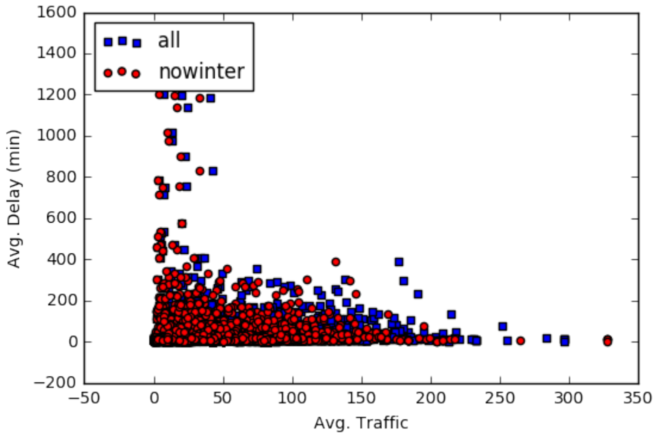


Fig. 1. Scatter-plot of avg. traffic vs. avg. delay. nowinter - stats excluding winter months

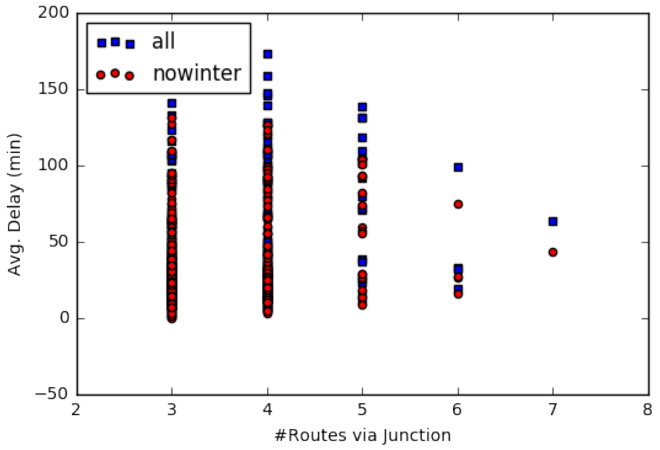
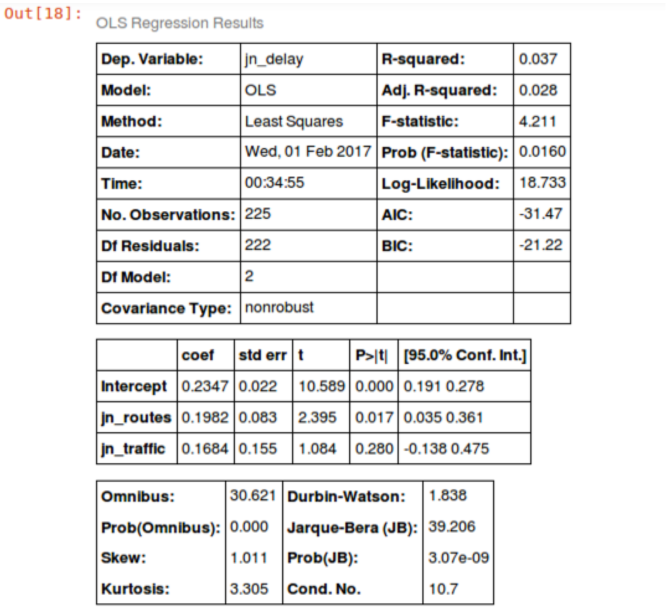


Fig. 2. Scatter-plot of no. of routes passing vs. avg. delay. nowinter - stats excluding winter months


Out[18]:

OLS Regression Results

Dep. Variable:	jn_delay	**R-squared:**	0.037
Model:	OLS	**Adj. R-squared:**	0.028
Method:	Least Squares	**F-statistic:**	4.211
Date:	Wed, 01 Feb 2017	**Prob (F-statistic):**	0.0160
Time:	00:34:55	**Log-Likelihood:**	18.733
No. Observations:	225	**AIC:**	-31.47
Df Residuals:	222	**BIC:**	-21.22
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.2347	0.022	10.589	0.000	0.191 0.278
jn_routes	0.1982	0.083	2.395	0.017	0.035 0.361
jn_traffic	0.1684	0.155	1.084	0.280	-0.138 0.475

Omnibus:	30.621	**Durbin-Watson:**	1.838
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39.206
Skew:	1.011	**Prob(JB):**	3.07e-09
Kurtosis:	3.305	**Cond. No.**	10.7

Fig. 3. Linear Regression Results