RAS Problem Solving Competition: Benchmark

Summary

Exploratory statistical research has been done to create the benchmark for this competition. Historical realization data was grouped together in a frequency table on basis of the delay in the initial and the delay final location. Predictions were made by taking the mean, median and trimean of each set. Significant differences between out-of-sample predictions were tested using a two-sample Z-test under the assumptions of the Central Limit Theorem. Data was split in different ways to improve performance. This way of prediction performed significantly better that the current algorithm, except with large delays.

General Principle

The research question that belonged to the creation of the benchmark is as follows:

“Is a statistical model that predicts delays of train routes based on historical delay information significantly better than the current prediction method?”

The data that was used in this research was the realization data from 2010, 2011 and 2012. To find patterns in the data, it was transformed into a frequency table. The row indicated the amount of delay at the current location (in minutes) and the column indicated the amount of delay at the location 20 minutes into the future. Trains that arrived more than two minutes early were not taken into consideration, resulting in that the row index runs from -2 up to \( n \). Here \( n \) is defined as the amount of delay where we put the cut-off point, all the delays larger than \( n \) will also be included in \( n \)th index. This was done as above a certain threshold, the amount of data points becomes really small as this does not happened that often. To ensure sufficient observations, these large delays are grouped together. In the end, we have \( n \) collections of delays, where the amount of elements in each collection is the sum over the frequencies of the second locations.

With the use of an estimator a prediction will be generator per collection. Three different estimators will be tested: the mean, the median and the trimean. The trimean is defined as a weighted average of the quartiles and the median:

\[
TM = \frac{Q_1 + 2Q_2 + Q_3}{4}
\]

Where \( Q_i \) denotes the i-th quartile.

Different methods exist for determining the first and third quartile. We chose to use the method of Hoaglin et al., where if two numbers are considered, the number that is closest to the median is chosen. This assures that the trimean is less sensitive to outliers.
Predictions will be rounded to whole minutes and will have a minimum of 0 to ensure we never predict an early departure.

**Measurement of Performance**

To test whether the performance of the algorithm is significantly better than the current algorithm, a margin is set in place to ensure a prediction does not need to be to the second exact. The performance of the estimators will be seen as the probability parameter of the Binomial distribution. A successful prediction will happen with probability $p_i$ (where $i \in$ estimators). For every $i$ the amount of successful predictions will be distributed as follows:

$$X_i \sim \text{Bin}(n, p_i)$$

Where $n$ is the amount of used observations.

For all combinations of estimators (including the current prediction method) we can test the following null hypotheses:

$$H_0 : p_i = p_j$$

We take $i$ and $j$ in such a way that $p_i \geq p_j$, so we can test the following one-sided alternative hypothesis:

$$H_1 : p_i > p_j$$

We test the hypotheses by making use of a two-sample Z-test under the assumptions of the Central Limit Theorem. If the Central Limit Theorem is violated, when $np_i < 5$ OR $n(1 - p_i) < 5$, the results are discarded. We assume that $\hat{p}_i - \hat{p}_j$ is asymptotically normally distributed with the following parameters:

$$\mu = \mathbb{E}[\hat{p}_i - \hat{p}_j] = \mathbb{E}[\hat{p}_i] - \mathbb{E}[\hat{p}_j]$$

$$= \frac{\mathbb{E}[X_i]}{n} - \frac{\mathbb{E}[X_j]}{n} = \frac{np_i}{n} - \frac{np_j}{n}$$

$$= p_i - p_j \overset{H_0}{=} 0$$

$$\sigma^2 = \mathbb{V}[\hat{p}_i - \hat{p}_j] = \mathbb{V}[\hat{p}_i] - \mathbb{V}[\hat{p}_j]$$

$$= \frac{\mathbb{V}[X_i]}{n^2} - \frac{\mathbb{V}[X_j]}{n^2} = \frac{np_i(1 - p_i) + np_j(1 - p_j)}{n^2}$$

$$= \frac{p_i(1 - p_i) + p_j(1 - p_j)}{n}$$

As we test under the null hypotheses, we won’t use $\hat{p}_i$ and $\hat{p}_j$ for $p_i$ and $p_j$, but we will make use of a pooled fraction:
\[ \hat{p} = \frac{X_i + X_j}{n_i + n_j} = \frac{X_i + X_j}{2n} = \frac{\hat{p}_i + \hat{p}_j}{2} \]

If we fill the pooled fraction in the variance equation above, we get the following sampling variance:

\[ s^2 = \frac{2\hat{p}(1 - \hat{p})}{n} \]

Which leaves us with the following test statistic of the two-sample Z-test:

\[ Z = \frac{\hat{p}_i - \hat{p}_j - \mu}{\sqrt{s^2}} = \frac{\sqrt{n}(\hat{p}_i - \hat{p}_j)}{\sqrt{2\hat{p}(1 - \hat{p})}} \]

**Split in Data**

To optimize performance, research was performed on how to best split the data. Reason behind this research was the fact that different hours, days, months, years might have different patterns in the data. We compared performance between data pooled together versus a single part of the data. A margin of two minutes was used to compare whether the two groups performed significantly different using the Z-test explained above. If more than 5% of the tests resulted in a significant difference, it was decided to not pool the data for that category together. The most important category comparisons were in this case whether it is rush hour or not, defined between 07:00 and 09:00 or between 16:00 and 18:00, week days versus weekend days, which season it is and what year the data is from. The comparisons within the categories all turned out be significantly different from each other.

**Results and Conclusion**

With the data split according to the significant differences described above, results show that using either the mean, median or trimean predictions are more accurate than the current algorithm for a delay not bigger than five minutes. Performance was expressed in terms of the percentage accurate predictions with a margin of two minutes. Unfortunately, for delays bigger than five minutes, the current algorithm still outperforms the estimators. The different estimators did differ in performance, the median and trimean both outperformed the mean. There was no significant difference found in performance between the median and trimean.

**Discussion**

A limitation of the performed research is that we work with one prediction per collection (which is the delay at the initial location). It could happen that some collections have two distinct distributions in their data, e.g. caused by the fact that sometimes a train can get stuck behind a slower train, but in other locations this does not happen. This means the estimators won’t capture the change in delay well. To distinguish between the different distributions, it might be beneficial to include the interactions with other trains. It could also be that different train routes are too different to be easily pooled together. Further research in this field is needed to answer these questions.