Analysis of delays on the New Jersey railway network

Azouz, Rami
Charif, Linah
Espadaler Clapés, Jasso
Fayed, Lynn

21 January 2020

École
polytechnique
fédérale
de Lausanne

Outline

1 Introduction

NJ Transit + Amtrak (NEC) Rail Performance dataset available on Kaggle

- 2nd largest railway network in the US in terms of number of passengers
- Commuting between NJ and NY

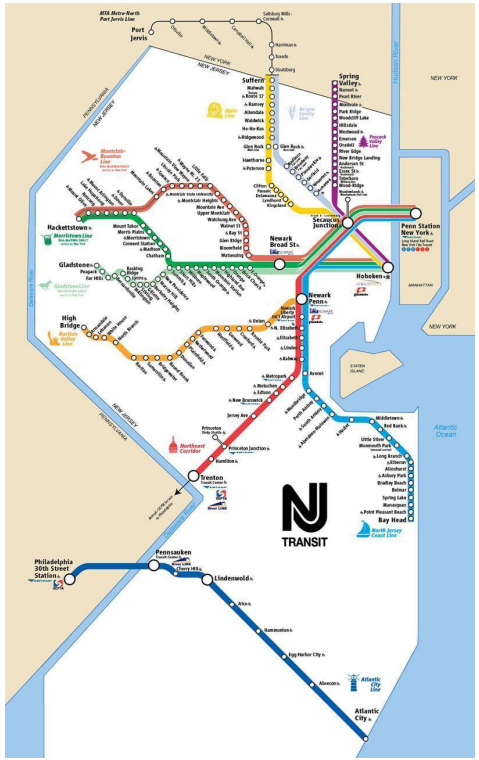
⇩

Evaluating delays and identifying their causes is essential

⇩

How are delays distributed in the network?

Can we predict them?



2 Preprocessing

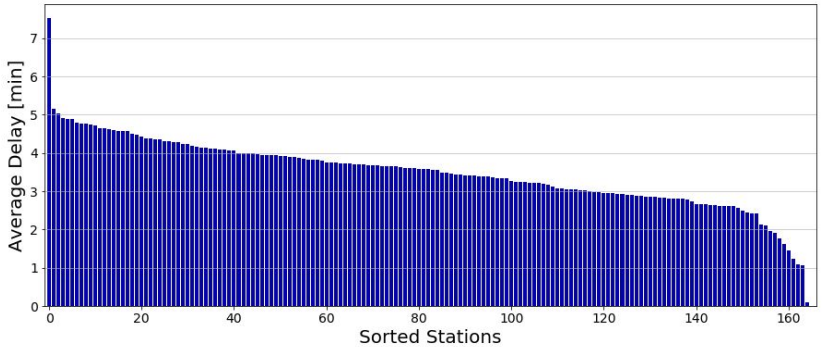
Dataset structure:

- Train trips for the NJ transit network are available at stop-level and with minute resolution
- Train trips from March 1, 2018 to April 30, 2019
- Only for March 2018:

Observations	Stations	Lines	Trains
243028	165	11	1319

How to evaluate delays?

- Average delay per station during the month of March 2018

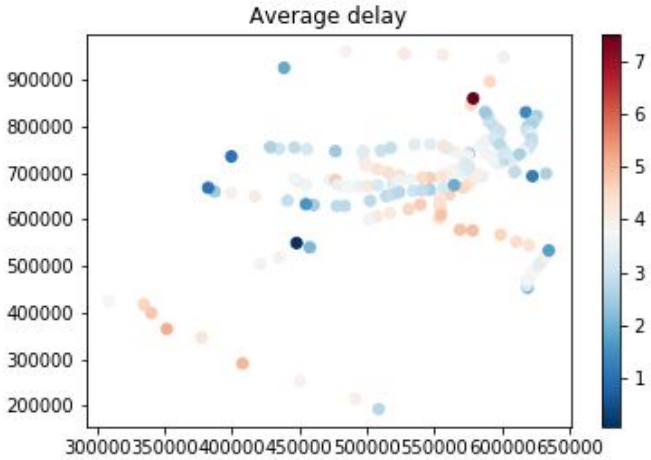


2 Preprocessing

Choice of metrics:

Objective: Find a scheme for data aggregation along each station

- Accentuation of node delays between stations
- Four different potential metrics have been attempted
- Cumulative sum and mean have shown to perform the best



Data filtering:

Objective: Noise elimination

→ Exclusively considering the weekdays and removing weekend delays

→ Removing remote stations with no direct interaction with the rest of the network

→ Filtering by inbound and outbound trains to take into account directional variation of delays

3 Clustering

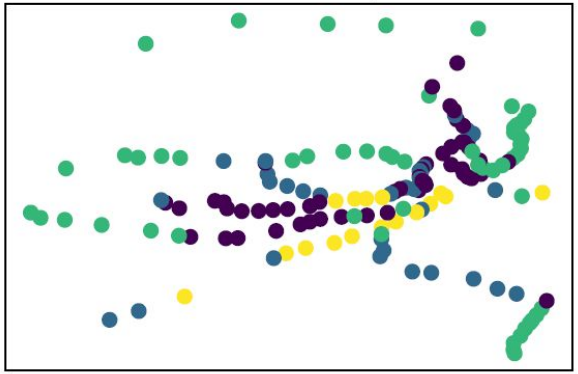
Spectral clustering:

- Similarity of cumulative delays or number of trains crossing each station.

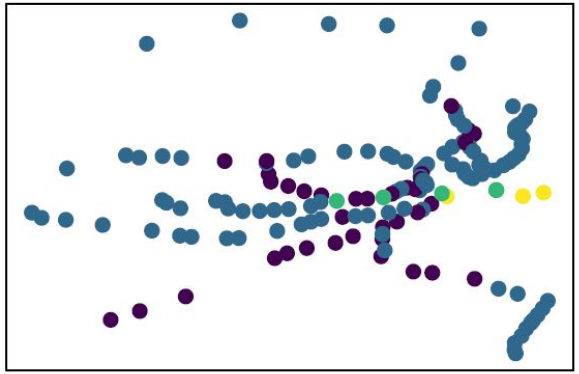
$$\omega_{ij} = \exp\{-|x_i - x_j|/2\sigma^2\}$$

- Difficult to tune σ and ε to get only one connected component: adjacency is not sparse.
- Misleading results from 4+ clusters.

Based on delays



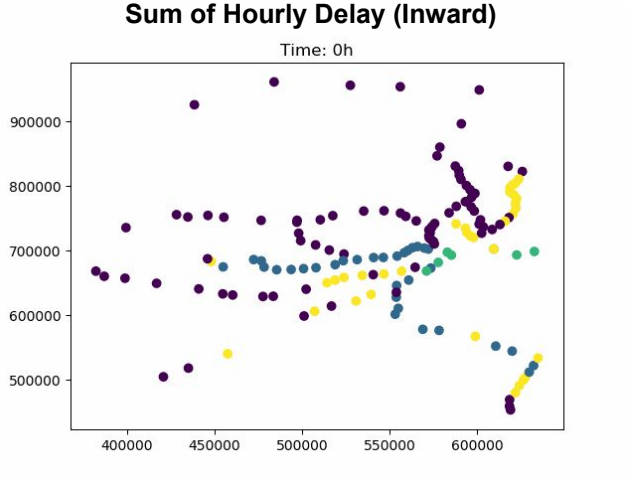
Based on no. of trains

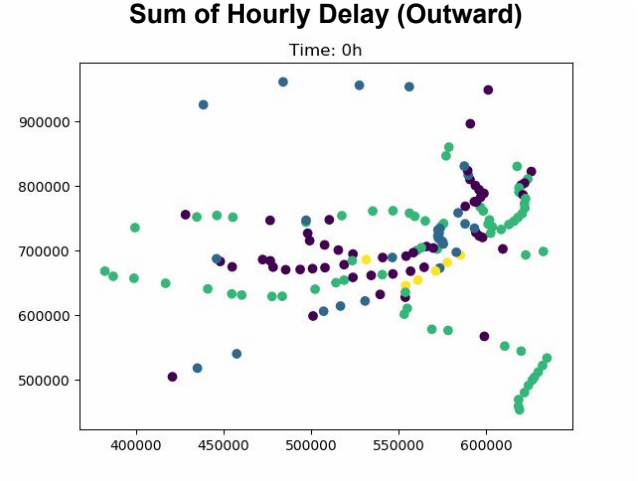


3 Clustering

k-means:

- Used to track down how clusters will vary over the day for inbound and outbound trains
- Helps the operator to identify problematic stations
- Gets insight on how delay propagates in the network
- More than 4 clusters lead to chaotic results



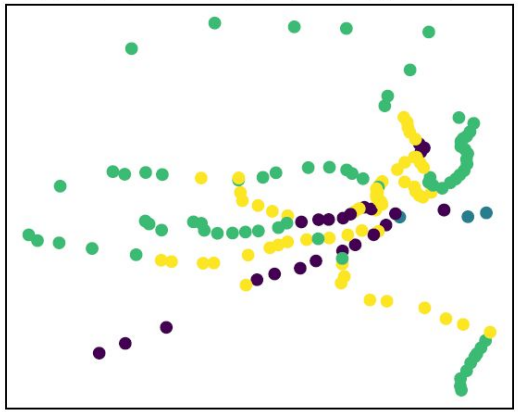


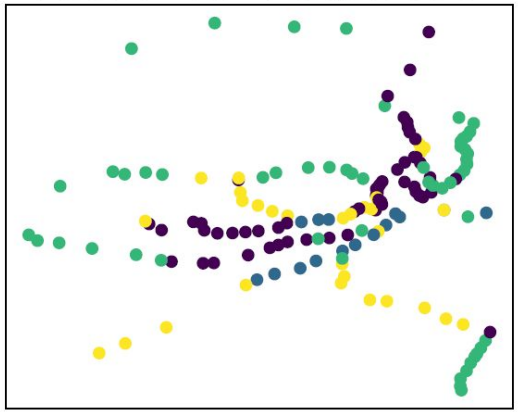
3 Clustering

k-means:

Initial Assumptions: Train frequencies are connected to cumulative delay levels in the network and distinction should be made between central and peripheral stations

- Possible to detect terminal stations for each graph
- Assumption validated with respect to spatial clustering of stations
- Similarities observed between clusters of the two graphs

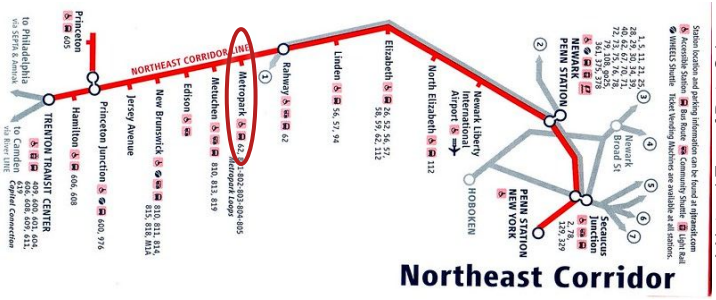




4 Prediction of delays

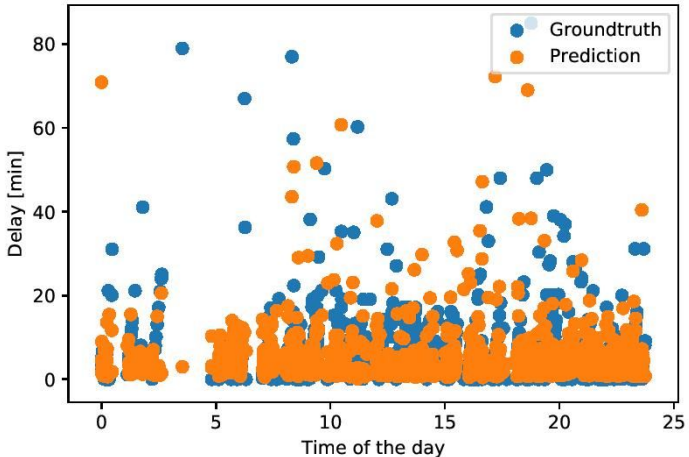
RNN with LSTM:

- **Preprocessing :** [#observations, #previous stations, #features]
  - 19403 observations
  - Features : time of arrival and delay for a given stop + the four previous stops



- **Model :**
  - Trained on one-month data for a single line *Northeast Corridor*, the features at all stops have been considered except for the stop chosen for the test set : *Metropark*
  - Prediction for the delays at *Metropark*

- **Results :**
  - Prediction of delay across the time of arrival at *Metropark* station during 24h
  - 20% accuracy
  - 10 min MSE



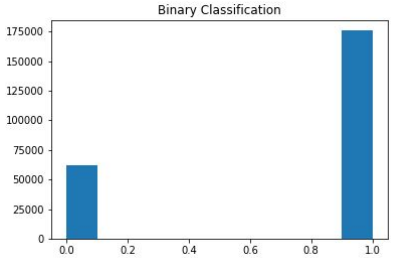
4 Prediction of delays

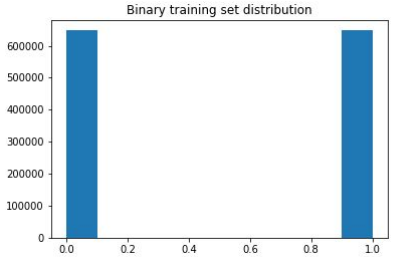
ANN Binary Classification:

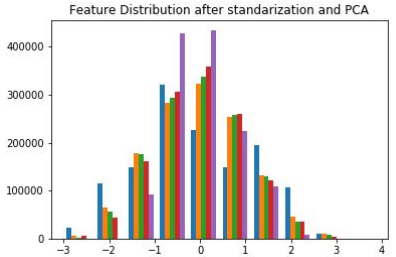
- **Preprocessing :**
  - 1038522 observations
  - Features : train_id, stop_sequence, line, day and time
  - Output: 0 if delay less than 1 min, 1 otherwise
  - Data standardization and PCA (d = 5)

- **Model :**
  - ANN:
    - 9 hidden layers of 30 nodes each
    - 2 dropout layers with rate 0.2
  - Training over one year, test on 1 month

- **Results :**
  - 62 % of accuracy on a balanced set
  - Difficult to improve prediction



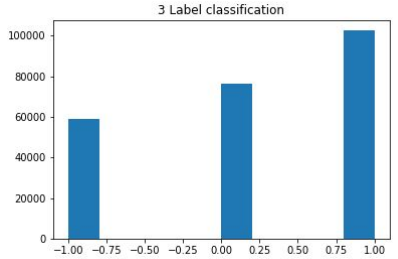


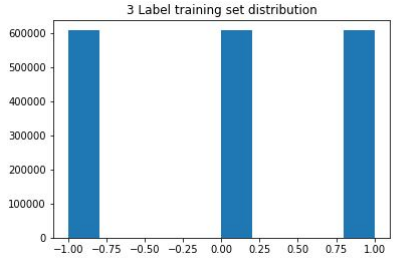


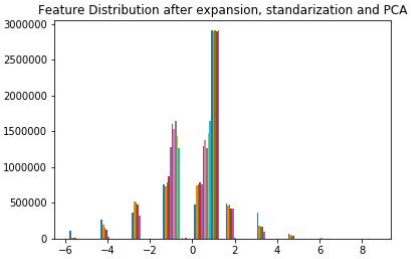
4 Prediction of delays

ANN Multi-label Classification:

- **Preprocessing :**
  - 1458562 observations
  - Features : train_id, stop_sequence, line, day and time
  - Output: -1 if delay less than 1 min, 0 if delay between 1 and 7 min and 1 otherwise
  - MinMax scaling and feature expansion with exp(X), sin(X) and cos(X)
  - Data standardization and PCA (d = 15)
- **Model :**
  - ANN:
    - 9 hidden layers of 30 nodes each
    - 2 dropout layers with rate 0.2
  - Training over one year, test on 1 month
- **Results :**
  - 41 % of accuracy on a balanced set
  - Difficult to improve prediction







5 Conclusion

Clustering:

- Correlation between node position, delay, and train frequency at each station is obvious for the central and peripheral stations
- Results are inconclusive for intermediate stations

RNN:

- Could possibly be a powerful tool
- Requires a lot of parameter tuning and training

ANN:

- Not the best tool to model the delay prediction problem
- A Graph based machine learning model maybe more adapted for our application

Analysis of delays on the New Jersey railway network

Azouz, Rami
Charif, Linah
Espadaler Clapés, Jasso
Fayed, Lynn

21 January 2020

École
polytechnique
fédérale
de Lausanne