Predicting and Clustering Station Vulnerability in Urban Networks

A Case Study for the Washington Metro

Menno Yap
Oded Cats

TU Delft – Smart Public Transport Lab
smartPTlab.TUDelft.nl
M.D.Yap@TUDelft.nl

July 9th, 2019
Study Relevance

- Study relevance:
  - PT disruptions: impact for PT passenger and operator
  - Most PT vulnerability studies focus on disruption *impacts*
  - Can incorrectly focus on severe, but rare disruptions

- Public transport vulnerability = frequency * impact:
  - Exposure to different disruption types at different stations
  - Limited availability of disruption log data over longer period
Objectives / Contribution

- Analyse disruption exposure characteristics based on empirical incident log data

- Develop a prediction model to predict future disruption exposure:
  - To different disruption types
  - For different stations
  - For different time periods of the day

- Obtain insights in susceptibility of different station types to disruptions:
  - To prioritise (type of) stations to consider for potential mitigation measures
Definitions

- Vulnerability $V = \text{exposure} \times \text{impact (pass-hours)}$
- Station criticality $c_s$: contribution of individual station to $V$

\[ E(c_s) = \sum_{t \in T} \sum_{d \in D} E(f_{d,t,s}) \times E(w_{d,t,s}) \]  

\[ V = \sum_{s \in S} E(c_s) / \sum_{t \in T} \sum_{s \in S} \sum_{s \in S} u_{t,s,s} \]  

station $s \in S$

time period $t \in T$
disruption type $d \in D$
disruption frequency $f$
disruption impact $w$
total passenger travel time $u$

- Definition of disruptions from incident log data:
  - Incident with either train delay or line delay $\geq 2$ minutes
  - Cleaning / interpreting log data to make data fit for purpose
Disruption Classification

Disruption category

Railcar
- Door malfunction
- Brake malfunction
- ATC malfunction
- Propulsion malfunction
- Other malfunction

Operations
- Action / error dispatcher
- Action / error train operator
- Station / signal overrun

Public
- Unattended item
- Trespassing incident
- Injured / sick / aggressive person

Infra
- ATC / power / track failure

Other
- Smoke / fire
- Collision / jumper / derailment
- Other / unknown

Introduction
Methodology
Case Study
Results
Conclusions
Disruption Exposure Prediction

- Prediction of disruption probability per type, station and time period:
  - Samples = |S| * |T|

- Supervised learning:
  - Logistic regression
  - Multilayer perceptron

- Randomised 5-fold cross validation

- Hyperparameter tuning: find optimal number of neurons for hidden layer MLP

- Log-loss (entropy loss) as performance metric
Disruption Exposure Clustering

• Unsupervised learning to cluster stations based on predicted disruption exposure

• Hierarchical agglomerative clustering:
  o Complete clustering
  o No pre-determined number of clusters

• Determine optimal number of clusters from dendrogram:
  o Input: matrix with shape $(|S|, |D|)$ consisting of values $E(f_{d,s})$
  o Distance matrix: $|D|$-dimensional Euclidean distance
  o Ward used as linkage-criterion
  o Average silhouette coefficient used to determine optimal $k$
Case Study: Washington D.C.

- 6 metro lines
- 95 metro stations
- Network length ≈190 km

Incident database WMATA:
- 13 months (Aug ‘17 – Sept ‘18)
- 7,263 incidents
- 5,835 distinguishable disruptions

- 13-months AFC + AVL data
Empirical Results: Type / Line

- 25% due to actions / errors dispatchers, terminal supervisors or interlocking supervisors
- 20% due to injured / sick / aggressive passengers
- 12% due to door malfunctioning
- 2/3 of disruptions is operation- or vehicle-related
Empirical Results: Spatial

- Start / terminal stations: highest exposure to disruptions
Model Estimation Results

- Feature vector: $|S| \times |T| = 95 \times 991 = 94,145$ samples and 34 columns (one-hot encoding)
- Target vector: 16 columns (binarised disruption classes)
- MLP: log-loss after hyperparameter tuning
- Optimal value: 29 neurons for hidden layer

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Log loss</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression (random 5-fold cross validation)</td>
<td>0.2817</td>
<td>0.6779</td>
</tr>
<tr>
<td>Multilayer perceptron classifier (random 5-fold cross validation)</td>
<td>0.2758 (-2.1%)</td>
<td>0.7107 (+4.8%)</td>
</tr>
</tbody>
</table>
Disruption Prediction Results

- High correlation (> .995) between observed and predicted disruption exposure per year
- Slight underestimation of predicted disruption frequency (8%)
Disruption Clustering Results

- Two clusters with terminals
- Cluster with transfer stations
- Cluster with all other stations
Conclusions

- Development of a model to predict future disruption exposure and to cluster stations accordingly

- Largest potential for disruption exposure reduction:
  - At start / terminal and transfer stations
  - Operations-related and vehicle-related disruptions

- Future work: integrate disruption impacts with predicted disruption exposure

Acknowledgment of WMATA for cooperation and data provision
TU Delft – Smart Public Transport Lab: smartptlab.tudelft.nl
Contact: M.D.Yap@TUDelft.nl