Predicting and clustering station vulnerability in urban networks

Menno Yap, Delft University of Technology, M.D.Yap@TUDelft.nl
Oded Cats, Delft University of Technology, O.Cats@TUDelft.nl

Extended abstract
Prepared for TransitData 2019: 5th International Workshop and Symposium: Research and applications on the use of passive data from public transport
Word count (excl. figures and references): 997

Introduction

Disruptions in public transport (PT) can have major implications for passengers and the public transport operator (e.g. *Hörcher et al. 2017* and *Yap et al. 2018*). The majority of scientific studies to PT vulnerability focus on the *impact* of a disruption, once a disruption occurs at a certain location in the network (e.g. *Derrible and Kennedy 2010*, *Cats and Jenelius 2015* and *Cats and Jenelius 2018*). PT vulnerability is however influenced by both the frequency and the impact of disruptions. Focusing solely on vulnerability in relation to disruption impacts can incorrectly put the emphasis on very severe, but very rare disruptions. Studies to PT disruption exposure are however limited (e.g. *Cats et al. 2016*). An important reason is often the lack of disruption log data, as data over a longer period of time is required given the relatively infrequent occurrence of disruptions.

The objective of our study is to develop a generic methodology to accurately predict and cluster vulnerability of different stations of a PT network, thereby incorporating both disruption exposure and impact based on the location-specific characteristics of the different stations. We apply our approach to the Washington metro network, using a 13-month database with logged disruptions received from the Washington Metropolitan Area Transit Authority.

Methodology

Definitions

We define *vulnerability* as the degree of susceptibility of a PT network to disruptions and the ability of PT network to cope with these disruptions (*Rodriguez-Nunez and Garcia-Palomares 2014*; *Oliveira et al. 2016*). Moving from a network level to individual elements, we define criticality as the degree an individual station contributes to vulnerability. For a given PT network we define each station of the total set $s \in S$, each disruption type $d \in D$, and each time period $t \in T$. When we define the disruption frequency f and the disruption impact w, the expected station criticality c for different disruption types is defined by (1), measured in passenger delay time. PT network vulnerability V can then be defined as the sum of station criticality, expressed as fraction of the total passenger travel time u as per (2).

$$E(c_s) = \sum_{t \in T} \sum_{d \in D} E\left(f_{d,t,s}\right) * E(w_{d,t,s}) \qquad (1)$$

$$V = \sum_{s \in S} E(c_s) \;/\; \sum_{t \in T} \sum_{s \in S} \sum_{s \in S} u_{t,s,s} \qquad (2)$$

Disruption classification

The large number of incident types in the provided disruption log data is classified into a small and distinctive number of disruption types based on their root cause. This results in a sufficient number of observations per disruption type and station to develop a vulnerability prediction model for. Consequently, all disruptions are classified in 15 different distinctive types $d \in D$ (Figure 1).

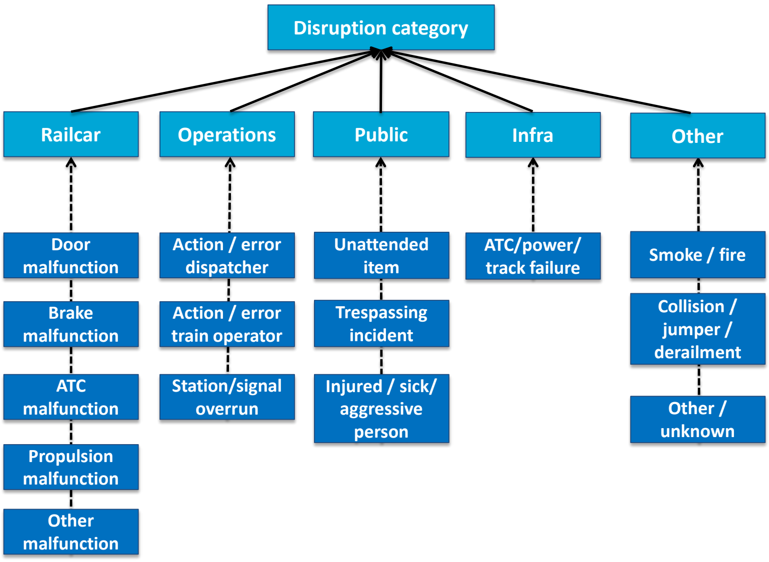

Fig. 1. Classification of disruptions.

Prediction of disruption exposure and impact

We develop two separate supervised learning models to predict exposure to, and impact of different disruptions $d \in D$ at stations $s \in S$ during each time period $t \in T$. This allows us to find linear and non-linear relations between presumed predictors and the exposure and impact of disruptions. Given the relative infrequent occurrence of disruptions, we use the probability of each disruption type as target for the prediction of disruption exposure. To this end, we test a logistic regression and multilayer perceptron (MLP) classifier. The total dataset is split into a 80% training set and 20% testing set, applied in a randomized 5-fold cross validation. Several location-specific station characteristics are identified as predictor in our machine learning model: day of the week, time period, season, the lines serving each station, whether the considered station is a terminal or transfer station, passenger volume - obtained from AFC data – and train frequency. Besides, the disruption exposure in the previous month is used as predictor auto correlating in time with the target variable (Figure 2). A comparable framework is currently being developed to predict disruption impact based on location-specific characteristics of the different stations. By multiplying the predicted disruption exposure and impact resulting from the two models, we quantify station vulnerability for each station.

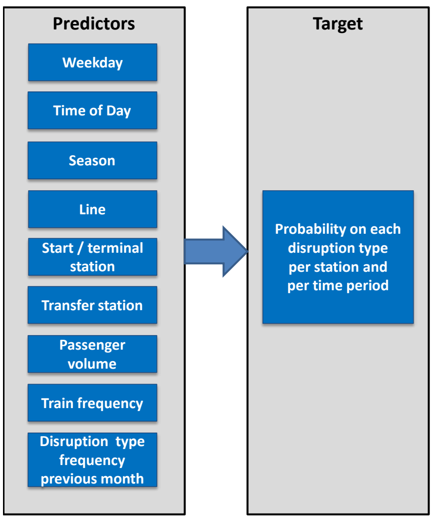

Fig. 2. Framework of prediction model for disruption exposure.

Clustering stations based on predicted vulnerability

To obtain insight in differences in susceptibility for different disruption types between stations, all stations are clustered based on disruption exposure, disruption impact, and the combined vulnerability. We use hierarchical agglomerative clustering as unsupervised learning approach, as we need to cluster all stations without a pre-defined number of clusters.

Results

Prediction of disruption exposure and impact

Using the provided incident log data, the spatial distribution of disruptions over the Washington metro network is visualized in Figure 3. As using a MLP classifier resulted in the lowest log-loss value, we present results based on this supervised learning model. Based on hyperparameter tuning we use 29 neurons for the intermediate layer of this model. From Figure 4 can be concluded that two third of all disruptions are operations-related (action / error dispatcher or train operator) or vehicle-related (primarily malfunctioning of doors and brakes). There is a very high correlation (>0.995) between our predicted numbers and observed values. On average the expected number of disruptions is underestimated by 8% using our model, indicating there is still room for further improvement.

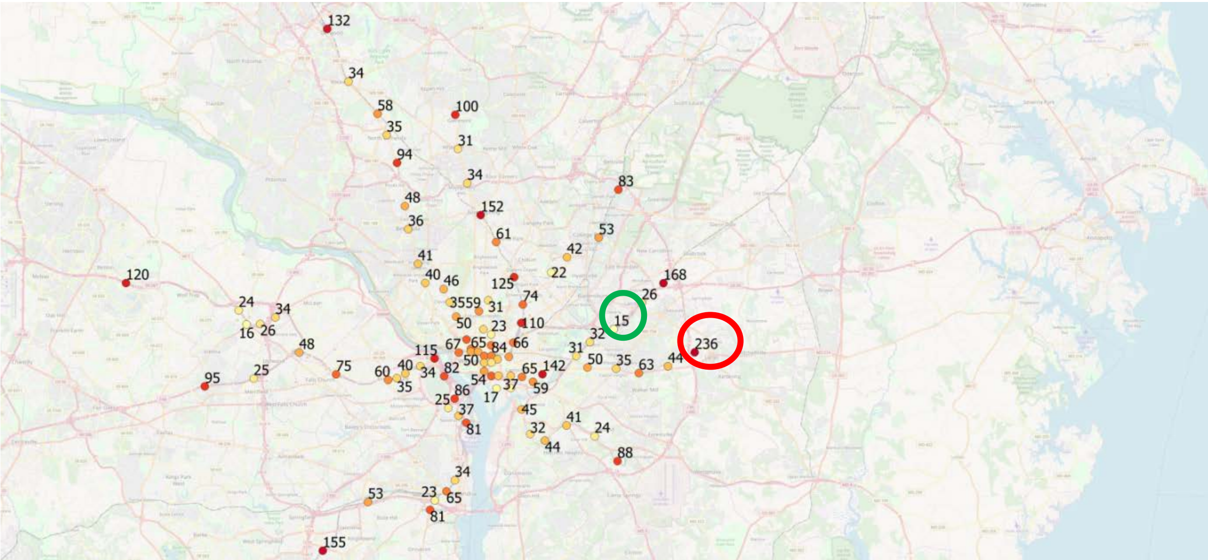

Fig. 3. Spatial distribution of disruptions.

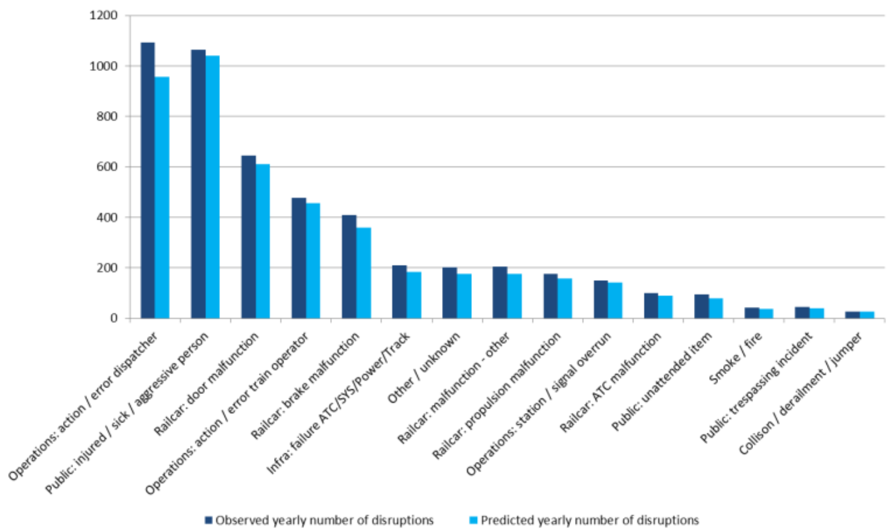

Fig. 4. Prediction of disruption exposure using a MLP classifier.

Station clustering based on disruption exposure and impact

Figure 5 presents the clustering results based on disruption exposure solely. Four clear groups of stations are found in this clustering. The first two clusters only consist of start/terminal stations, which show to be substantially more susceptible to disruptions than other stations. All transfer stations of our case study network are grouped into a separate cluster, indicating a distinctive exposure pattern related to the particular characteristics of transfer stations. All other intermediate, non-terminal and non-transfer stops are grouped together in one cluster as being least susceptible to disruptions. A similar clustering is currently performed based on disruption impact, and based on the product of exposure and impact.

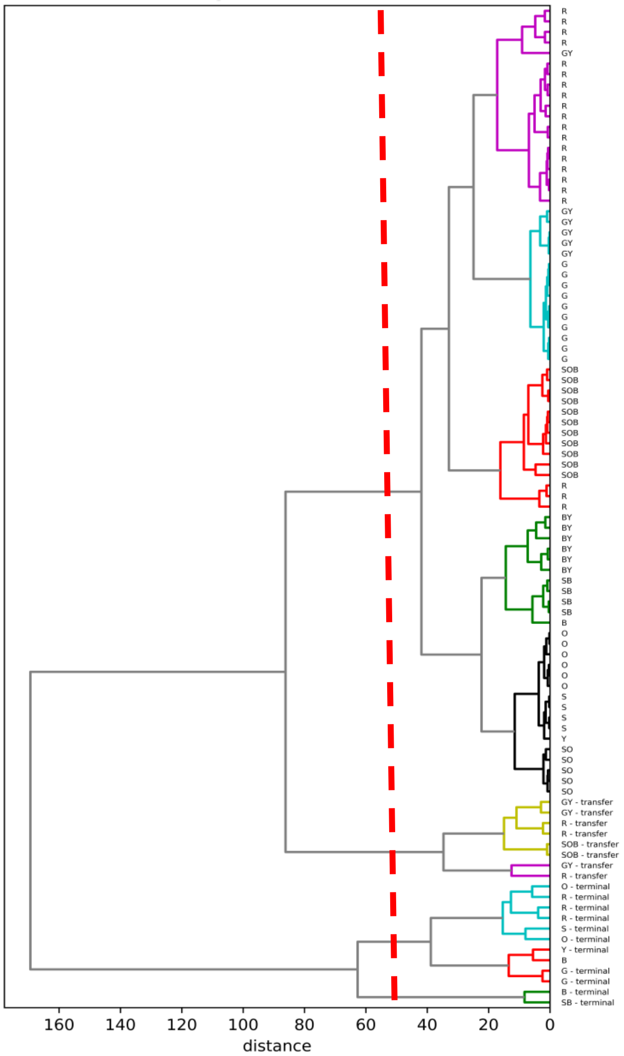

Fig. 5. Dendrogram with resulting clustering of metro stations based on disruption exposure.

Policy implications

Our study results provide PT authorities and operators insight to the type and location of disruptions which contribute most to total network vulnerability, based on predicted disruption exposure and impact. This supports them in prioritizing what type of disruptions at what location to focus on, to potentially achieve the largest increase in network robustness. Our further work will focus on the development of a prediction model for disruption impacts, which will be integrated with the disruption exposure prediction results.

References

Cats, O. and E. Jenelius, “Planning for the unexpected: the value of reserve capacity for public transport network robustness,” Transportation Research Part A, vol. 81, pp. 47-61, 2015.

Cats, O., and E. Jenelius, “Beyond a complete failure: the impact of partial capacity degradation on public transport network vulnerability,” Transportmetrica A: Transport Dynamics, vol. 6, pp. 77-96, 2018.

Cats, O., M.D. Yap, and N. van Oort, “Exposing the role of exposure: Public transport network risk analysis,” Transportation Research Part A, vol. 88, pp. 1-14, 2016.

Derrible, S. and C. Kennedy, “The complexity and robustness of metro networks.” Physica A, vol. 389, pp. 3678-3691, 2010.

Hörcher, D., D.J. Graham, and R.J. Anderson, “Crowding cost estimation with large scale smart card and vehicle location data,” Transportation Research Part B, vol. 95, pp. 105-125, 2017.

Oliveira, E.L., L. Silva Portugal, and W. Porto Junior, “Indicators of reliability and robustness: Similaries and differences in ranking links of a complex road system,” Transportation Research Part A, vol. 88, pp. 195-208, 2016.

Rodriguez-Nunez, E., and J.C. Garcia-Palomares, “Measuring the vulnerability of public transport networks,” Journal of Transport Geography, vol.35, pp. 50-63, 2014.

Yap, M.D., S. Nijenstein, and N. van Oort, “Improving predictions of public transport usage during disturbances based on smart card data,” Transport Policy, vol. 61, pp. 84-95, 2018.