Predicting disruptions and their passenger delay impacts for public transport stops

Yap, Menno; Cats, Oded

DOI
10.1007/s11116-020-10109-9

Publication date
2020

Document Version
Final published version

Published in
Transportation

Transportation (2021) 48:1703–1731
https://doi.org/10.1007/s11116-020-10109-9

Predicting disruptions and their passenger delay impacts for public transport stops

Menno Yap[1] · **Oded Cats**[1]

Published online: 28 April 2020

Abstract
Disruptions in public transport can have major implications for passengers and service providers. Our study objective is to develop a generic approach to predict how often different disruption types occur at different stations of a public transport network, and to predict the impact related to these disruptions as measured in terms of passenger delays. We propose a supervised learning approach to perform these predictions, as this allows for predictions for individual stations for each time period, without the requirement of having sufficient empirical disruption observations available for each location and time period. This approach also enables a fast prediction of disruption impacts for a large number of disruption instances, hence addressing the computational challenges that rise when typical public transport assignment or simulation models would be used for real-world public transport networks. To improve transferability of our study results, we cluster stations based on their contribution to network vulnerability using unsupervised learning. This supports public transport agencies to apply the appropriate type of measure aimed to reduce disruptions or to mitigate disruption impacts for each station type. Applied to the Washington metro network, we predict a yearly passenger delay of 5.9 million hours for the total metro network. Based on the clustering, five different types of station are distinguished. Stations with high train frequencies and high passenger volumes located at central trunk sections of the network show to be most critical, along with start/terminal and transfer stations. Intermediate stations located at branches of a line are least critical.

Keywords Disruptions · Machine learning · Passenger delay · Public transport · Vulnerability

Introduction

Relevance

Disruptions in public transport (PT) can have major implications for passengers and service provider. Disruptions can increase passengers' nominal travel time, due to additional

✉ Menno Yap
M.D.Yap@TUDelft.nl

[1] Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands

waiting time, in-vehicle time or transfers. Furthermore, passengers potentially experience higher crowding levels on alternative services, resulting in a more negatively perceived in-vehicle time (Hörcher et al. 2017; Tirachini et al. 2017; Yap et al. 2018a). Disruptions can also imply costs for the service provider, due to overtime payments to personnel, possible fare reimbursement for delayed passengers, and in the case of contractual agreements between service provider and authority resulting in fines. In the long term, disruptions can result in a loss of revenue if ridership levels decrease because of (perceived) unreliability of the PT system. It is thus in the interest of passengers and service provider to examine and assess the frequency, location and passenger delay impact of different disruption types occurring at each public transport station or link. An accurate prediction of the occurrence and impact of disruptions supports PT authorities and service providers in prioritising the locations and disruption types for which they should devise measures to reduce disruptions or their impacts.

Definitions and scope

For the remainder of this paper, we first introduce several definitions used throughout this work. We apply a definition of *vulnerability,* which is obtained by combining definitions from Rodriguez-Nunez and Garcia-Palomares (2014) and Oliveira et al. (2016), with *robustness* being its antonym. Vulnerability is defined as the degree of susceptibility of a PT network to disruptions and the ability of the PT network to cope with these disruptions. This definition highlights the two components vulnerability consists of: *exposure*, the degree to which a PT system is exposed to disruptions, and the *impact* once a disruption occurs. Moving from a network level to individual elements, we define *criticality* as the degree to which an individual element of a PT system—such as a PT node or link—contributes to vulnerability. Criticality again refers to both disruption exposure and impact: it considers both *weakness*, the degree of disruption exposure for an individual stop or link, and *importance,* the impact of disruptions occurring at a stop or link (Cats et al. 2016). The most critical nodes or links thus contribute most to PT vulnerability in terms of the product of their weakness and importance.

We consider both recurrent and non-recurrent PT disruptions in our study. Recurrent PT disruptions, such as a vehicle door malfunctioning or a delayed departure from the terminal, occur relatively frequently whilst the impact is generally limited. To the contrary, non-recurrent PT disruptions—such as a faulty train, signal failure or vehicle derailment—are relatively rare, but often have larger impacts once they occur. In this study, we focus on the impact of identifiable, distinctive disruptions. The impact of normal stochasticity of the PT system, for example caused by variability in train running times, is not considered. In addition, we do not consider extreme events as natural disasters or terror attacks in this research. These events differ substantially from typical PT disruptions in terms of frequency, location and impact, that a bespoke research approach is necessary. Furthermore, we focus on unplanned disruptions: planned disruptions, for example related to scheduled track maintenance works, fall outside the scope of this work.

State-of-the-art

Empirical data can contain information about the frequency with which different disruptions occurred, or about the disruption impact on passenger delays. However, to be able to study disruption frequencies and impacts of different disruption types for individual

elements of a PT network, only using empirical data is typically insufficient. For illustration purposes, let us consider a medium-sized PT network consisting of 100 stops, where our aim for each stop is to predict disruption frequencies and impacts for 20 different disruption types, for five different time periods of the day and week, separately for each season. This would require empirically deriving disruption frequencies and impacts for $100 \times 20 \times 5 \times 4 = 40{,}000$ instances. Consequentially, this would require sufficient empirical observations for each of these 40,000 instances to fit a probability density function for, to use empirical data reliably to predict future disruption frequencies or impacts. In practice, this means there will be insufficient empirical data available from past disruptions to use directly for future disruption occurrences and impacts for each of these instances. Therefore, some kind of prediction model becomes necessary to predict disruption frequency and impact for each individual PT network element.

In the field of transport vulnerability analysis, different approaches are applied to predict disruption impacts: *full scan* computation methods and methods using pre-selection indicators (Knoop et al. 2012). Full scan methods predict the disruption impact of each disruption type, at each location of a PT network. In a wider context, approaches to predict disruption impacts are broadly classified as scenario-based, strategy-based, simulation-based or using mathematical modelling (Murray et al. 2008). For transportation networks, generally a static, dynamic or simulation-based transport assignment model is used for this purpose. For example, in the context of highway networks, Jenelius (2007) uses a traffic simulation model where each link of the network is blocked, whilst Knoop et al. (2008) also incorporate dynamic spillback effects of blocked links. Full scan methods result in impact predictions for each individual network component, hence allowing all stops or links being ranked according to their contribution to network vulnerability. However, these methods are computationally prohibitive for larger networks and are typically only feasible to apply for smaller or case study networks. Instead, pre-selection methods apply indicators which result in a short-list of locations where disruption impacts are expected to be most severe. Full disruption impacts are only modelled or simulated for this selection of locations. For example, Tampère et al. (2007) assess the expected criticality of road network links based on multiple indicators, such as the incident impact factor. Other road network vulnerability indicators used in literature are the Network Robustness Index (Scott et al. 2006) and the Modified Network Robustness Index (Sullivan et al. 2010), which approximate the impact of a full or partial link blockage on the network performance, respectively. Bell (2003) and Zhang et al. (2010) both adopt a game theoretical approach to quantify indicators for network vulnerability. To assess vulnerability of PT networks, Derrible and Kennedy (2010) propose a robustness indicator which calculates the number of available paths in the event of a disruption, based on a graph representation of 33 metro networks worldwide. Cats et al. (2016) compare a passenger betweenness centrality measure as proposed by Cats and Jenelius (2014) and a passenger-exposure measure as PT vulnerability indicator. Aforementioned studies adopt either a node-based or link-based vulnerability approach, whilst some studies consider the vulnerability impacts of joint node and link disruptions (see for example Dinh and Thai 2014). The disadvantage of pre-selection approaches however is that there is no guarantee the largest impacts occur at these selected locations. This means there is no certainty whether the most critical nodes or links of a network are correctly identified. Additionally, these approaches do not allow for a comparison of disruption impacts between all individual network elements, as disruption impacts are only quantified for selected elements. The abovementioned state-of-the-art illustrates that existing methods are insufficient to predict the passenger delay impacts from disruptions for each individual PT station or link for medium- or large-sized, real-world PT networks.

Several studies focus on predicting disruption impacts, once a disruption occurs. For example, studies quantify PT disruption impacts (Cats and Jenelius 2015), the value of spare capacity in a PT network (Cats and Jenelius 2014), or the impact of partial rather than complete track closures (Cats and Jenelius 2018). Corman et al. (2014) evaluate the robustness of railway timetables once a disruption occurs. However, focusing solely on disruption impacts without considering disruption frequencies can incorrectly put the emphasis on locations where very severe yet very rare disruptions occur. Predicting how often different locations in a PT network are exposed to different disruptions is a relatively understudied topic. There has been a vast amount of work towards predicting incident frequencies for road networks in the field of traffic safety. Whereas initial road traffic research primarily used descriptive and aggregate models to predict accident probabilities (e.g. Stone and Broughton 2003; Lord et al. 2005), more recent research has moved towards using disaggregate, predictive models (e.g. Zou and Yue 2017). For PT networks, the use of disaggregate models remains limited. An important reason is often a lack of good quality disruption log data, as data over a longer period of time is required given the relatively infrequent occurrence of disruptions. In Cats et al. (2016) and Yap et al. (2018b), a database consisting of logged disruptions on a PT network for a period of 2.5 year was used to fit statistical models for disruption frequencies on a network level. In these studies, relatively simple predictors such as the number of trains or train-kilometres were used to translate the network-wide number of disruptions to expected disruption exposure per station or link. This implies that location-specific characteristics—such as the type of stock serving a station, the passenger load or the geographical area where a station is located—are not considered, while these are believed to be important when predicting disruption exposure. In Tonnelier et al. (2018) a data-driven method is developed to detect individual atypical events in PT networks using anomaly detection. However, this method does not explicitly provide what type of disruption at which location initiated this anomaly, making it difficult to formulate policy recommendations concerning how to tackle PT vulnerability. This means that currently no adequate disaggregate models have been developed to predict disruption frequencies for individual PT stops or links.

Research approach and contribution

Our study objective can be summarised as the development of a generic methodology to predict disruptions and their passenger delay impacts accurately for different disruption types, for individual stations of a real-world PT network, thereby incorporating the specific characteristics of the different stations. This implies we develop a disaggregate modelling approach to predict disruption frequencies and to predict the passenger delay impacts of each disruption. We propose a supervised learning approach to perform these predictions, as this allows for the prediction of disruptions at individual stations for each time period, without the requirement of having sufficient empirical disruption observations available for each location and time period. This approach also enables a fast prediction of disruption impacts for a large number of disruption instances, hence addressing the computational challenges that rise when typical PT assignment or simulation models would be used for real-world PT networks.

To improve transferability of our study results, we cluster stations based on their contribution to PT vulnerability using unsupervised learning. Besides predicting disruptions and their impacts for a specific PT network, this provides PT authorities and service providers insight in the different station types that can be distinguished based on their contribution

to network vulnerability. For example, for policy purposes train stations in the Netherlands are grouped into six categories based on function and passenger volumes (Geurs et al. 2016). Our research results in a natural clustering of stations in a similar way, specifically based on vulnerability. Hence, this supports PT agencies to apply the appropriate type of measure aimed to reduce disruptions or to mitigate disruption impacts for each station type. Our research contribution is therefore defined as follows.

Scientific contributions

- Development of a method to predict disruptions and their passenger delay impacts for individual public transport stations, incorporating the specific characteristics of each station;
- Development of prediction models which predict disruptions and their impacts based on a non-exhaustive empirical disruption dataset within acceptable computation times.

Practical contributions

- To provide PT agencies with predicted disruption impacts for each individual station on their network, for each distinguished time period and disruption type, supporting them to prioritise locations where to put mitigation measures in place;
- Identification of different groups of public transport stations with different disruption exposure and impact characteristics, enabling PT agencies to devise appropriate measures to tackle vulnerability for different station types.

The remainder of this paper is structured as follows. The Methodology section explains the methodology, whilst the Case Study section introduces our case study network. We discuss results in the Results and Discussion section, followed by the Conclusions section.

Methodology

In this section we discuss the proposed methodology to predict disruption exposure and impact at different PT stations, and to cluster stations accordingly. First, we introduce the proposed modelling framework. Then, we explain our supervised learning model to predict disruptions, followed by the model for disruption impact predictions. At last, we discuss our station clustering approach. First, we introduce sets, indices and variables as used throughout the paper in Table 1.

Modelling framework

For a given PT network, let us define each station $s \in S$, with $|S|$ being the total number of stations in the considered network. Each disruption type is defined by d, with D indicating the total set. Each distinguished time period is indicated by $t \in T$. When we define the disruption frequency f and the disruption impact w, the predicted station criticality $\breve{c}$ in its simplest form is defined by Eq. (1). To obtain station criticality, the predicted frequency of each disruption (expressed in disruptions per year) at station s is multiplied by the predicted impact, and then summed over all disruption types and time periods considered per year. In our study, we predict the total passenger delay hours $\breve{w}_{d,t,s}$ as metric for disruption impact. It should be

Table 1 List with sets, indices and variables

Sets and indices	
d, D	Disruption type, set of disruption types
e, E	Vertex of graph G, set of vertices
i	Origin stop (vertex) of graph G
j	Destination stop (vertex) of graph G
s, S	Public transport stop, set of stops
t, T	Time period, set of time periods
v, V	Edge of graph G, set of edges
y, Y	Label in classification model, set of labels
Variables	
c	Station criticality
f	Disruption frequency
g	Percentage demand for which no simple path remains available during a disruption
h	Passenger-weighted travel time increase (hours)
jt	Journey time (hours)
l	Shortest path length (hours)
n	Number of shortest paths
p	Disruption probability
q	Passenger demand
t	Time (hours)
v	Network vulnerability (delay hours per year)
w	Passenger delay (hours)
x	Dummy variable

noted that disruption impacts are generally wider than passenger delays only. Passengers' perceived travel times often increase as well, whilst PT service providers might face rescheduling costs (e.g. personnel overtime payment) or passenger reimbursement costs. In this study we however only consider the nominal travel time impact a disruption has inflicted on passengers. This means station criticality is expressed in yearly passenger delay hours. PT network vulnerability V then equals the sum of the predicted station criticality (Eq. 2), and expresses the predicted yearly passenger delay hours for the total PT network of interest. For the sake of simplicity, the basic impact calculation as shown in Eq. (1) does not show interdependencies between different disruptions occurring simultaneously on the considered PT network, as this can result in interaction effects affecting the disruption impact. The integrated modelling framework to calculate $\breve{c}_s$ and $\breve{v}$ is shown in Fig. 1. It shows the supervised learning models used to predict disruptions and passenger delay impacts, as well as the unsupervised learning model applied to categorise different stations. This modelling framework is explained further in the remainder of this section.

$$\breve{c}_s = \sum_{t \in T} \sum_{d \in D} \breve{f}_{d,t,s} \times \breve{w}_{d,t,s} \tag{1}$$

$$\breve{v} = \sum_{s \in S} \breve{c}_s \tag{2}$$

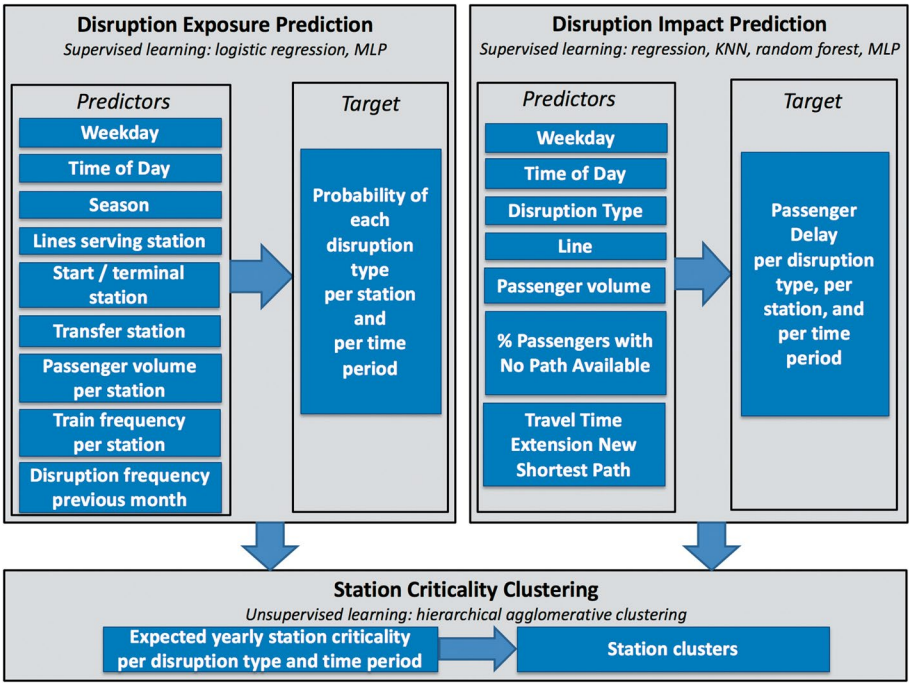

Fig. 1 Modelling framework

For our proposed modelling framework, the following empirical data sources are required as input:

- Disruption log data, containing data for each PT disruption which occurred on the PT network within the considered time interval. As a minimum, for each disruption the start time, location and line of occurrence need to be logged, as well as the disruption type or a disruption description. Disruption end time is desirable though not mandatory for our method. This type of data is usually available at the PT authority or service provider, based on logged incident notifications from train drivers, station operators, control room staff, police and the general public.
- Individual passenger demand data from Automated Fare Collection (AFC) systems, which consists of the time and location of the first boarding and final alighting station of each individual passenger journey. This allows calculation of the realised journey time for each passenger.
- Scheduled journey times between each boarding and alighting station for each distinguished time period or day of the week, allowing a comparison between scheduled and realised passenger journey times. This data can be obtained from journey planners or can be provided by the PT service provider.
- Timetable data from GTFS or Automated Vehicle Location (AVL) systems (typically open data), which contains the planned number of PT trips for each route, during each time period and day of the week.

Disruption exposure prediction

In this study we adopt a supervised learning approach to predict exposure to 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 disruption predictors and the exposure to disruptions with short computation times. As each disruption type occurs relatively infrequently at a specific station and within a specific time period, our study objective here implies predicting the occurrence of relatively rare events. For that reason we do not use $\breve{f}_{d,t,s}$ as our target, as a model always predicting zero for $\breve{f}_{d,t,s}$ would still result in a low *MSE-score* and high average *F1-score* due to the overrepresentation of samples with zero disruptions, without providing any useful information for predicting disruption exposure. Neither applying different weights for false positive and false negative predictions, nor applying a technique to correct the dataset imbalance such as a *Synthetic Minority Oversampling Technique* (SMOTE) did sufficiently improve the quality of disruption predictions. Thus, we use instead the probability $\breve{p}_{d,t,s}$ of each disruption type occurring within each considered time period (e.g. each AM, PM, Inter Peak and Evening period for each day of the year) as target for the prediction. $\breve{f}_{d,t,s}$ is calculated by multiplying the predicted probabilities by the number of time periods $|T|$. The number of samples in our model thus equals $|S| \times |T|$. To predict disruption probabilities we apply a classification algorithm, which calculates disruption probabilities for each $d \in D$ and then assigns each sample to one of the disruption categories d or to the category *no disruption* based on the highest probability. The dimension of the target vector therefore equals $(|S| \times |T|, 1)$, where column values can take $|D| + 1$ different values. In our case, this value equals 0 if no disruption is predicted to occur in the considered time period, and ranges between 1 and $|D|$ depending on which disruption type is predicted to occur within the time period. By dummy coding this target vector, a matrix with dimensions $(|S| \times |T|, |D| + 1)$ results which contains the predicted probabilities for each disruption type.

We identify several general and location-specific station characteristics as predictors in our machine learning model (Fig. 1, upper left). We first use the general predictors *Weekday, Time of Day* and *Season*. *Weekday* equals 1 if the time period is during a weekday, and 0 if during a weekend. *Time of Day* considers if the time period is during the peak (7–10 AM or 3–7 PM: only during weekdays), daytime off-peak (weekdays: hours outside peak until 7 PM; weekend: all hours until 7 PM) or evening (hours after 7 PM). The aim of these predictors is particularly to capture the possible influence of differences in mixture of passenger types and travel purposes between peak, off-peak, evenings and weekends on disruption probabilities. The predictor *Seasons* aims to capture differences in disruption probabilities for different seasons. One can think of potentially more vehicle defects due to leaves in autumn, or more passenger-related incidents due to slippery surfaces in winter. Additionally, we identify several station-specific predictors. *Lines* refers to the different metro lines serving each station, as different stock types on different lines potentially influence especially railcar-related disruption probabilities. A possible difference in state and age of infrastructure between different lines can also play a role here. One-hot encoding is applied for the categorical predictors *Time of Day, Seasons* and *Lines*, resulting in separate binary predictors for each category. If a station is served by multiple lines, for example being part of a trunk section, the binary predictor equals 1 for each of these lines. Two separate binary predictors *Start station* and *Transfer station* are added, being equal to 1 if the station is a start/terminal or a metro-to-metro transfer station, respectively. It is expected that the occurrence of some disruptions is related to a station being a start/terminal, as

problems such as a malfunctioning train or a late/absent train driver often arise here. It is hypothesised that transfer stations might be more susceptible to disruptions due to more complex infrastructure (such as switches) and large passenger transfer volumes. *Passenger volume* refers to the number of boarding plus alighting passengers for each station and time period, based on Automated Fare Collection (AFC) data for an average day. This predictor is added to capture primarily passenger-related disruption probabilities. *Train frequency* equals the scheduled number of trains serving a stop during each time period and day of the week. This predictor is calculated based on timetable data, and aims to capture railcar-related disruption probabilities. *Disruption frequency previous month* is an auto regressor with respect to the number of disruptions that occurred during a certain time period, weekday/weekend at the considered train station in the previous month, for each disruption type separately. This predictor assumes disruption data of the previous month is available to predict disruption exposure in the next month. In total, $|D|$ separate predictors are used for this predictor, for each disruption type $d \in D$. Values for predictors *Passenger volume, Train frequency* and *Disruption frequency previous month* are all normalised between 0 and 1, so that all predictors use the same range.

Given our target to predict the probability of different disruption types in a certain time period at a certain station, we test two different machine learning algorithms suitable for this purpose: logistic regression and a Multilayer Perceptron (MLP) classifier, a class of feedforward artificial neural networks. The total dataset is split into an 80% training set and 20% testing set, applied in a randomised fivefold cross validation. Applying a higher tenfold cross validation did not significantly improve prediction accuracy. As we are predicting the probability of different disruption types, we use log-loss (or cross entropy loss: Eq. 3) as accuracy metric. The log-loss function calculates the negative log-likelihood of the true label y, given the predicted probability that a sample equals this true label $\breve{y}$. In addition, we calculate the F1-score as accuracy metric. By calculating this F1-score globally by counting total true positives, false negatives and false positives, we account for label imbalance as existing in our dataset.

$$-logp_{y|\breve{y}} = -\left(y \times \log\left(p_{\breve{y}}\right) + (1-y) \times \log\left(1-p_{\breve{y}}\right)\right) \tag{3}$$

Disruption impact prediction

For the prediction of passenger impacts of disruptions, we also apply a supervised learning approach. To quantify passenger delays, we compare the scheduled and realised passenger journey times at the considered PT network. For each journey between a given origin station $i \in S$ and destination station $j \in S$, the realised journey time $jt_{t,ij}$ is obtained from AFC data per time period t. Depending on whether a tap in only or tap in/tap out AFC system is in place, the destination of an AFC transaction is directly available or needs to be inferred using a destination inference algorithm (e.g. Munizaga and Palma 2012). Besides, transfer inference might be required to connect AFC transactions to journeys, if transfers are made which require an intermediate AFC transaction (e.g. Gordon et al. 2013; Yap et al. 2017). The scheduled journey time for time $t \in T$ is calculated from the timetable. In-vehicle times and station walking times are assumed to be deterministic. The maximum scheduled journey time $\breve{jt}^{max}_{t,ij}$ assumes the passenger wait time is equal to the planned headway at that time (i.e. in case a passenger has just missed a train), whilst the minimum scheduled journey time $\breve{jt}^{min}_{t,ij}$ assumes a passenger can board the PT vehicle directly (no waiting time).

The expected scheduled journey time $E\left(\breve{jt}_{t,ij}\right)$ is then calculated as the average between $\breve{jt}^{min}_{t,ij}$ and $\breve{jt}^{max}_{t,ij}$. Equation (5) shows the passenger delay calculation applied in our study, using dummy variable x_1 as defined in Eq. (4). A journey in time period t is considered delayed if the realised journey time exceeds the maximum scheduled journey time. To prevent underestimating passenger delays in that case, the delay is calculated as the difference between realised and expected scheduled journey time (expressed in minutes) and multiplied by demand $q_{t,ij}$.

$$x_{1,ij} = \begin{cases} x_{1,ij} = 1 \;\; if \; jt_{t,ij} > \breve{jt}^{max}_{t,ij} \\ x_{1,ij} = 0 \;\; if \; jt_{t,ij} \leq \breve{jt}^{max}_{t,ij} \end{cases} \tag{4}$$

$$w_t = \sum_{i \in S} \sum_{j \in S} \left[jt_{t,ij} - E\left(\breve{jt}_{t,ij}\right) \right] \cdot q_{t,ij} \cdot x_{1,ij} \tag{5}$$

The structure of the machine learning model is shown in Fig. 1 (upper right). The passenger delay $\breve{w}$ resulting from disruptions is used as target. It should be noted that this delay cannot be attributed directly to a certain disruption d, as several disruptions can occur spatially and/or temporally close to each other. As disruption end times are not always provided in disruption log data, the disruption duration cannot always be determined. However, even if the end time of a disruption would be known, knock-on effects on passenger delays can persist for up to six times longer than the duration of the initial cause (Malandri et al. 2018). Once a disruption is resolved, there is typically recovery time required to reschedule PT trips and personnel before the origin timetable is restored. Hence, in any case the disruption log data does not provide information when the passenger delay impact for passengers ended. To mitigate this problem, our model is being trained using a rolling horizon where the total passenger delay w from time hour t up to 2 h later $\left[t, t_{+2}\right]$ is used as target, as function of the considered disruption which started during t together with all other disruptions which started during this time window $\left[t, t_{+2}\right]$. This approach implies we consider disruption impacts up to three hours after the moment the disruption occurred. Although this can theoretically underestimate the impact of large disruptions somewhat, the majority of the disruptions on a PT network are typically relatively minor. Therefore, this time horizon is deemed reasonable to capture the complete disruption impacts for the vast majority of all disruptions. If the impact of disruptions which started at t would vanish before t_{+2}, the calculated passenger delay during t_{+2} is expected to be (close to) zero as well, meaning there is no penalty for adopting a relatively long time horizon for smaller disruptions. If a disruption end time would be available from the log data, a more accurate time period could be determined in the rolling horizon. For example, the end of the horizon could be set equal to the logged disruption end time, plus a time period reflecting recovery time as function of the logged disruption duration. We apply the final trained model to a new test data set where only one disruption per t and s occurs, to predict the pure impact of each disruption separately.

As generic predictors for disruption impact, we use predictors *Time of Day* and *Weekday* (weekday, Saturday or Sunday). For the predictor *Disruption type,* we apply one-hot encoding for all disruption types $d \in D$, where a disruption type is coded as 1 in case this disruption has occurred within the time window $\left[t, t_{+2}\right]$. The PT line is

also used as predictor. Each line on which a disruption occurred in this same time window $[t, t_{+2}]$ is coded as 1; other lines are coded zero. As it is expected that total passenger delay depends on the total passenger volume using the network at the considered time period, we also use the total demand q_t starting a journey in this time interval as predictor. As disruption impacts can propagate over the total PT network, the total demand summed over all origin–destination stations is used here. We use the *Percentage affected demand for which no (simple) path remains available* in case a disruption blocks services to/from station s as a station-specific predictor. The higher the percentage passengers for which no alternative routes remain available in case of a disruption, the larger the passenger delay for the affected passengers one might expect. To quantify this predictor g_t we first calculate the affected demand q_t^a. We represent the total PT network as directional graph $G(V, E)$ with each vertex $v \in V$ representing a stop and each edge $e \in E$ representing a direct PT connection between stops. For each OD pair we calculate the length of the shortest path (expressed in minutes) l_{ij} and the number of simple paths (without cycles) n_{ij} for the undisrupted scenario. We define the affected demand as the demand between OD pairs for which the shortest path length increases or for which no simple paths remain available if all disrupted stations s^d where a disruption occurred in the time window $[t, t_{+2}]$ are removed from $G(V, E)$ (Eq. 8). Based on this, g_t can be calculated as value ranging between 0 and 1 (Eq. 9). For all affected demand for which at least one simple path remains available, the *Expected detour time* on the network (expressed in minutes) can be used as an additional station-specific predictor for the full passenger delay impact. To this end, the increased length of the shortest path can be computed, so that the passenger-weighted average travel time extension $\Delta\bar{h}$ can be quantified as an additional predictor for passenger delays (Eq. 10). Equations (6) and (7) introduce the required dummy variables $x_{2,ij}$ and $x_{3,ij}$.

$$x_{2,ij} = \begin{cases} x_{2,ij} = 1 & if \;\; l_{ij}^d > l_{ij} \\ x_{2,ij} = 0 & if \;\; l_{ij}^d \leq l_{ij} \end{cases} \tag{6}$$

$$x_{3,ij} = \begin{cases} x_{3,ij} = 1 & if \; n_{ij}^d = 0 \\ x_{3,ij} = 0 & if \; n_{ij}^d > 0 \end{cases} \tag{7}$$

$$q_t^a = \sum_{i \in S} \sum_{j \in S} \left[q_{t,ij} \cdot \max\left(x_{2,ij}, x_{3,ij}\right) \right] \tag{8}$$

$$g_t = \frac{\sum_{i \in S} \sum_{j \in S} q_{t,ij} \cdot x_{3,ij}}{q_t^a} \tag{9}$$

$$\Delta\bar{h}_t = \frac{\sum_{i \in S} \sum_{j \in S} \left[q_{t,ij} \cdot \left(l_{ij}^d - l_{ij} \right) \cdot x_{2,ij} \right]}{q_t^a \cdot \left(1 - g_t\right)} \tag{10}$$

We test different supervised learning regression models to predict passenger delays. We apply a simple linear regression model as baseline, and compare these results with a K-Nearest Neighbours (KNN), Random Forest and Multilayer Perceptron (MLP) regressor. For all these regression models, we apply a randomised fivefold cross validation

and use the *RMSE* (root-mean-squared error) as performance metric. The total number of samples of our models equals the number of disruptions in our database.

Clustering station criticality

The output of the final models to predict disruption exposure and impact is used as input to cluster PT stations based on their expected criticality. The final disruption exposure model is applied to predict disruption probabilities for each disruption type d for one complete year, whilst the final disruption impact model is used to predict the impact for each disruption type at each station s separately for each time period. Multiplication using Eq. (1) results in the expected yearly criticality per disruption type, station and time period, expressed in yearly passenger delay hours. We apply an unsupervised learning method to cluster stations based on this expected criticality (Fig. 1, lower part). This provides insight in differences in susceptibility for different disruption types between stations, and shows clusters of stations with similar disruption exposure and impact patterns.

As our aim is to cluster all stations $s \in S$ without outliers, and no number of clusters k is known a priori, we apply hierarchical agglomerative clustering. Input for the clustering is a matrix consisting of values $\breve{c}_{s,d,t}$ with dimensions $(|S|, |D| \times |T|)$, which results from our supervised learning models. The distance matrix is determined by calculating the $|D| \times |T|$ -dimensional Euclidean distance between all points. Ward is used as linkage criterion during the clustering, thereby minimising the within-cluster variance. We use the cophenetic correlation coefficient to assess the degree to which the clustering reflects the input data. The optimal number of clusters k is determined based on visual inspection of the dendrogram and maximising the average silhouette coefficient. The silhouette coefficient for each sample is calculated by taking the difference between the Euclidean distance to the nearest cluster this sample is not part of, and the intra-cluster distance. This difference is then divided by the maximum value of these two. The average silhouette coefficient is obtained by calculating this for all $|S|$ stations.

Case study

Case study network

We apply our proposed methodology to the Washington D.C. metro network as case study. The Washington Metro, administered by WMATA, consists of six lines indicated by different colours: the Red line (R), Green line (G), Yellow line (Y), Blue line (B), Orange line (O) and Silver line (S) (Fig. 2). The total length of the metro network is about 190 km. During AM and PM peak hours, the Red line runs 15 trains per hour (tph), of which every other train is a short-turning service to Silver Spring. The other lines run 7.5 tph during peak hours. During daytime off-peak periods, all lines run 5 tph. The Blue, Orange and Silver line share a substantial part of their routes between Rosslyn and Stadium-Armory. The joint frequency on this trunk section equals 22.5 tph during peak hours. At the time of consideration, 95 different metro stations are operational, thus $|S|=95$. We predict disruption probabilities for all distinguished time periods (peak (only for weekdays), daytime off-peak and evening) for a full year. Every week thus consists of 19 time periods (3 time periods for weekdays and 2 time periods for weekend days). Hence, for a complete year $|T|$ equals

Fig. 2 WMATA metro network (Map obtained from WMATA: https://www.wmata.com/schedules/maps/upload/2019-System-Map.pdf)

991. For our case study network, the total number of samples for the disruption prediction model equals $|S| * |T| = 94{,}145$.

Input data sources

One of the important data sources used as input for our method is incident log data. A 13-month incident database for the Washington metro network is provided by WMATA, which initially consisted of 21,868 records covering all reported incidents from August 1st 2017 to August 31st 2018. The attributes of each record are shown in Table 2. This shows that each record consist of a start time, incident location, line, train id, disruption category and description. Besides, the minutes of initial train delay (delays for an individual train)

Table 2 Example incident log data

ID	Start Time	Line	Train	Stop	Type	Description	Train delay	Line delay	Initial incident
11	16-08-17 8:30	Blue	419	C07	AIRL	Air leak	5	5	11
12	23-08-17 9:13	Red	231	A11	PUBL	Sick customer	3	0	12

and line delay (delay for the entire line) as a result of an incident are indicated. This reflects only the initial delay a certain incident had on the train or line involved, and does not contain any information about the possibly (wider) passenger delay impact following this initial delay due to spill-over effects. The column *Initial incident* indicates if an incident is the result of another incident. The same number in the *Initial incident* column indicates that two logged incidents are related to each other. As the end time of disruptions is not logged for our case study, we use the time window $[t, t_{+2}]$ as rolling horizon in the disruption impact prediction, with t being the hour when the disruption starts (see the Methodology section). We use all disruptions in the 12-month period from September 1st 2017 to August 31st 2018 as input for our disruption exposure prediction model. Disruption data for August 2017 is used to quantify values for the auto regressor predictor *Disruption frequency previous month* (see the Methodology section) for disruption probabilities in September 2017.

Incident log data of PT systems is generally not primarily intended for vulnerability analysis purposes or to draw policy recommendations from. Instead, this is usually filled out during the real-time control process in the control room when recovering train services. This also entails there might be only a limited degree of consistency in the description and classification of incident notifications, as it strongly depends on manual actions from controllers whose main priority is solving the incident. This was also the case for the Washington data set provided to us. As a result, it is important to reassure the incident database is fit for our study purpose, for which we perform two data processing steps: (a) deriving disruptions from incidents, and (b) classifying disruptions.

First, we derive disruptions from incidents in the log file, as this database also contains incidents which did not result in a disruption. For example, a driver not able to perform its duty due to sickness is reported in the incident database, even if a stand-by driver took over the shift without any delays. For our case study, we define a disruption as any incident where either the train delay or line delay is 2 min or more. Incidents with both the train and line delay being smaller than 2 min are regarded as regular service variability. Additionally, when multiple incidents in the database are related to the same incident, only the initial incident is kept. Other delays can be considered a consequence of this initial incident, rather than separate incidents. When applying our disruption definition, 4263 distinguishable disruptions remain in the 12-month period from September 2017 to August 2018.

Second, disruptions are classified into a selected number of distinctive disruption types. In the provided database, 114 different disruption types are logged. When considering the distribution over different stations $s \in S$ and time periods $t \in T$, there would be an insufficient number of observations per station and time period in the database to develop a prediction model for. Besides, in some cases different definitions were used for the same or very similar disruption types, due to differences in classification by different controllers. For example, in the used database a train car motor overload is indicated by both disruption type *MOLD* ('motor overload') and *MOLF* ('flashing motor overload'). In these cases, one consistent disruption type is attributed to both disruptions. In some cases, the disruption types in the database did not reflect the root disruption cause. As an illustration, one can find an incident registered as *ONEC* ('operational necessity') with the description 'late dispatch due to door not closing'. In this case, the root cause is a door malfunctioning, resulting in an operational action from the control room. In a manual exercise, all disruptions in the database are classified based on their root cause following their description. Consequently, all disruptions are classified into 15 different distinctive types $d \in D$, which occur frequently enough to be able to develop a prediction model for. The distinguished disruption types are visualised by the dark-blue rectangles in Fig. 3. As can be seen in the

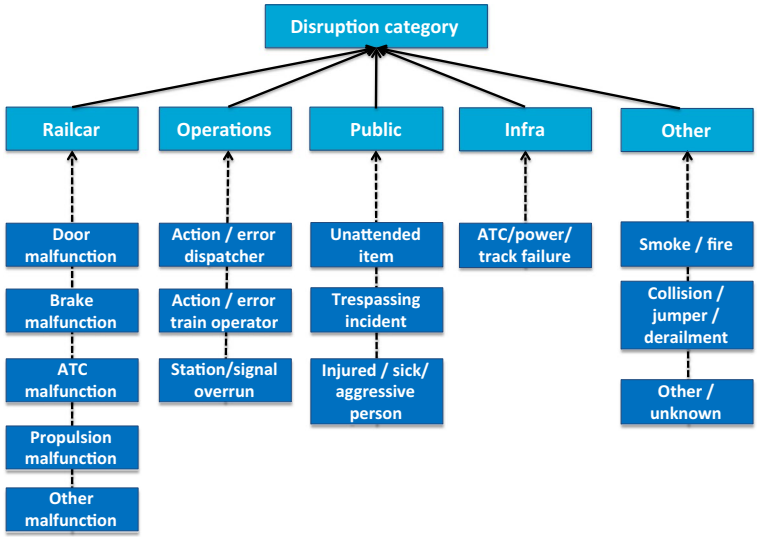

Fig. 3 Disruption classification

light-blue rectangles, these disruption types are classified into five main categories *railcar-related, operations-related, public-related, infrastructure-related* and *other* disruptions. The category with railcar-related disruptions for example consists of *door malfunctioning, brake malfunctioning, ATC malfunctioning, propulsion malfunctioning* and *other* disruption types. Similarly, public-related disruptions are categorised as *left unattended item, trespassing incident* and *injured/sick/aggressive passenger.* Weather-related disruptions are not considered a separate category in our study, as the impact of certain weather types influences the frequency of one or multiple of the 15 distinguished disruption types. For example, slippery platforms during winter might increase the number of injured passengers, whereas snow can affect the frequency of different railcar-related and infrastructure-related disruptions. As *season* is used as one of the predictors in the disruption prediction model (Fig. 1), it partially accounts for the impact of different weather types during different seasons on the disruption frequencies. In our study, we did not have access to detailed weather data (such as amount of precipitation, or temperatures for different areas and time periods). In the event that this data would be available, more detailed predictors (e.g. *mm rainfall per time and area*) could be used instead of the more generic predictor *season*.

Additional to disruption log data, timetable data about train frequencies and scheduled passenger journey times is provided by WMATA. Besides, individual AFC transactions of each journey made on the metro network in September, November and December 2017, as well as January, February and March 2018 were also available for this study. The Washington metro network is a closed system, where passengers are required to tap in and tap out at gatelines at the stations. For metro-to-metro transfers typically no intermediate tap out and subsequent tap in is required. This means that our case study data directly consists of the journey start time and end time for the metro network, so that no destination or transfer inference was required. Given the availability of 6 months AFC data, our disruption impact prediction model—for which AFC data is required as input for the predictors—is trained based on this 6-month data set. In this 6-month period, 2179 disruptions can be distinguished from the data set after applying the abovementioned data processing steps. This is

in contrast with the disruption exposure model, which is trained based on a 12-month disruption log data set. As we apply a fivefold cross validation, in each of the five folds 80% of this data is used for model training, whilst the remaining 20% is used for model testing purposes.

Empirical disruption characteristics

Figure 4 shows the spatial distribution of disruptions over the Washington metro network. The empirical values show that the weakest stations, being most susceptible to disruptions, can generally be found in the central area of the network where train frequencies and passenger volumes are highest, and at start/terminal stations. The least weak stations are typically intermediate stations (non-terminal and non-transfer stations) at the line branches, often served by one line only. Largo Town Center (red circle in Fig. 4) suffered from most disruptions in the observed 12-month period (160 disruptions). Figure 5 presents the relative frequency of the 15 different disruption types, categorised into the five main categories as set out in Fig. 3. It can be seen that vehicle-related disruptions contribute most to the total number of disruptions (45%). In total, vehicle-related and passenger-related disruptions are responsibe for more than 70% of all disruptions. Infrastructure-related disruptions only have a relatively small share in the total number of disruptions. From the individual disruption types d, the most frequently occurring types are injured/sick/aggressive passengers (23%) and vehicle door malfunctioning (15%).

Model specification

We use our developed model to predict the probability a certain disruption type occurs at each station during each time period (peak, daytime off-peak and evening) for one full year, applied to the Washington case study network. Based on the number of predictors and one-hot encoding, the final feature matrix for our exposure prediction model consists

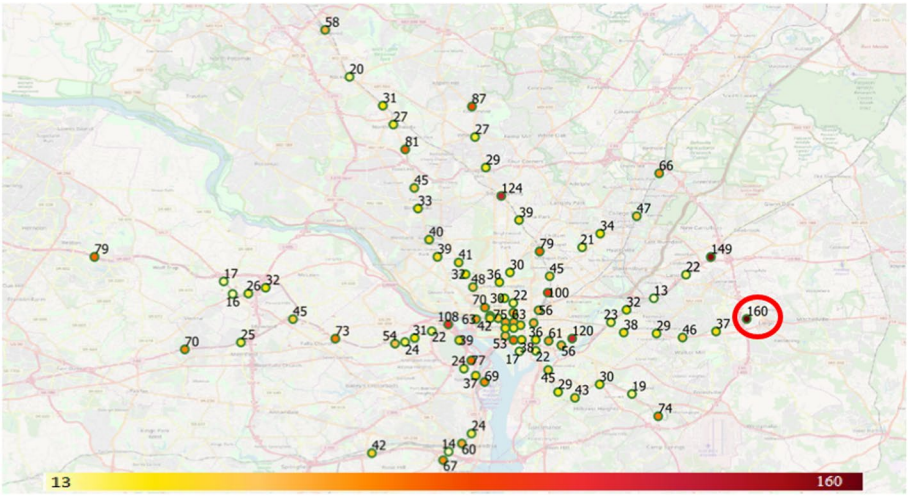

Fig. 4 Spatial distribution of yearly number of disruptions

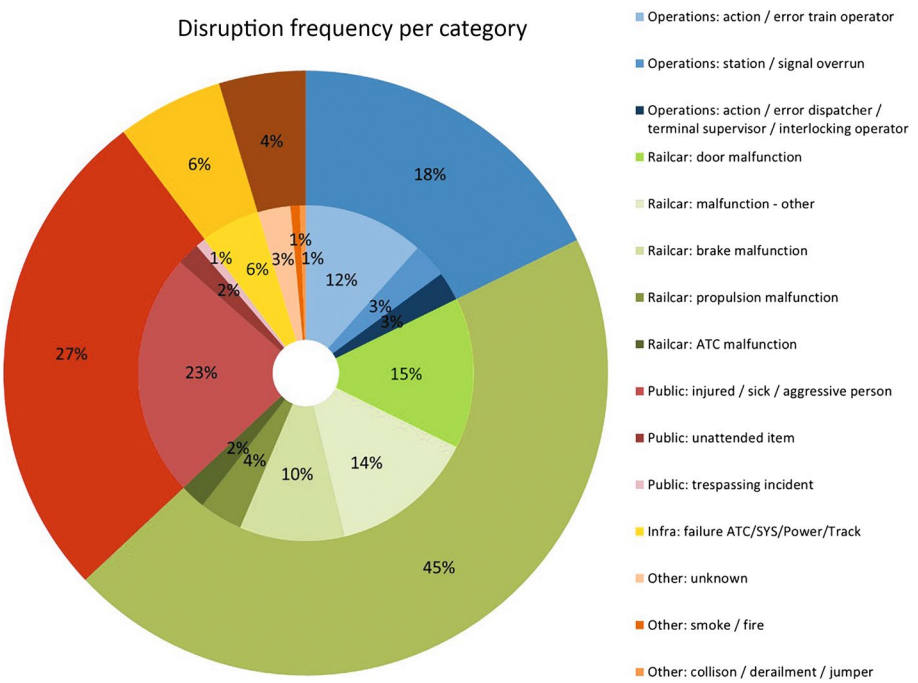

Fig. 5 Relative distribution of distinguished disruption types for 12-month period Sept.'17–Aug.'18. The outer circle reflects the share of each of the 5 main disruption categories distinguished in light blue in Fig. 3; the inner circle reflects the share of each of the 15 distinguished disruption types as reflected in dark blue in Fig. 3

of (991 distinguished time periods per year × 95 stations) 94,145 samples and 34 columns. The dimension of the target vector is (94,145; 1), respectively (94,145; 16) when dummy coded into the 16 disruption classes (15 disruption types plus no disruption). The *scikit-learn* library of Python is used to execute the machine learning models (Pedregosa et al. 2011). For the logistic regression model we perform a multiclass regression with a maximum of 200 iterations. *Sag* is used as solver method, as this is fast for relatively large datasets (Schmidt et al. 2017). For the MLP classifier one hidden layer is used. Furthermore, *Adam* is used as solver method, with the number of iterations being capped at 200. *Adam* is an adaptive learning rate optimisation algorithm, which allows a model to learn faster and converge earlier, resulting in better model performance (Kingma and Ba, 2015). A logistic sigmoid function is used as activation function for the hidden layer. The number of neurons of the hidden layer is determined by hyperparameter tuning: for all number of neurons between the number of neurons of the input layer and output layer the log loss score (Eq. 3) is calculated, thereby selecting the number of neurons for the hidden layer which minimises this value. The optimal number of neurons of the hidden layer for the MLP classifier is therefore sought between 16 and 34 neurons. The computation time to predict disruption probabilities for 1 year for the medium-sized Washington metro network (95 stations) is for both models less than 1 min on a regular PC.

The final feature matrix of the passenger delay prediction model consists of 2179 samples (disruptions) and 30 columns (7 one-hot encoded predictors). In the KNN algorithm,

we test K-values ranging between 1 and 30 during hyperparameter tuning. For the Random Forest model, we test the number of estimators between 100 and 1000 with step size 100 for a model which uses the total number of features as maximum feature number, and for a model using the square-root of the number of features as maximum feature number. For the MLP model, we specify a maximum of 10,000 iterations, use *l-bfgs* as solver and apply a logistic activation function (Byrd et al. 1995). During the hyperparameter tuning we test the number of neurons of the hidden layer between the number of neurons of the output layer (1) and input layer (30). Computation times for these four models to predict disruption impacts range between 1 and 10 min on a regular PC.

Results and discussion

In this section, we first discuss the model estimation and validation results. Then, we discuss the prediction results, followed by clustering results.

Model estimation and validation

Table 3 provides an overview of the model specification and performance results for the developed disruption exposure and disruption impact prediction models. Regarding the two tested disrupted exposure prediction models, it can be seen that the log-loss score and F1-scores are similar for the logistic regression and MLP classifier. The log-loss score of 0.268 can be considered reasonably close to 0, whilst the F1-score of 0.958 indicates a satisfactory model performance. For this case study we decide to proceed with the results

Table 3 Model estimation results

Disruption exposure model	Model specifications	Log-loss/F1 score
Logistic regression classifier *(Random fivefold cross validation)*	Max iterations = 200 Solver = sag	0.268/0.958
Multilayer Perceptron classifier *(Random fivefold cross validation)*	Max iterations = 200 Solver = adam Activation function = logistic Neurons hidden layer = 30	0.268/0.958
Disruption impact model	**Model specifications**	**RMSE (R^2) score**
Simple linear regressor *(Random fivefold cross validation)*	With intercept	2,536,946 (−551)
Simple linear regressor *(Random fivefold cross validation)*	Without intercept	1,575,072 (−292)
K-Nearest Neighbours regressor *(Random fivefold cross validation)*	Number of neighbours = 26	81,950 (0.57)
Random Forest regressor *(Random fivefold cross validation)*	Number of estimators = 200 Max number of features = features	64,722 (0.74)
Random Forest regressor *(Random fivefold cross validation)*	Number of estimators = 900 Max number of features = sqrt(features)	65,533 (0.73)
Multilayer Perceptron regressor *(Random fivefold cross validation)*	Max iterations = 10,000 Solver = lbfgs Activation function = logistic Neurons hidden layer = 25	115,971 (0.18)

from the MLP classifier, as this model can potentially capture more complex relations between predictors and target. For passenger delay predictions, different machine learning models are compared to a simple linear regression model as baseline. One can conclude that all machine learning models outperform the linear regression model substantially, reducing the RMSE by 95–97%. This indicates the linear regression model is not suitable to capture the complex relation between the predictors and target. When comparing the different machine learning models, especially the Random Forest and KNN regression models result in lower RMSE scores and reasonably high R^2 scores. The Random Forest model, with the total number of features as maximum number of features and using 200 estimators, results is the lowest RMSE score and highest R^2 score of 0.74. We therefore use this model for our final passenger delay predictions.

For model validation purposes of the disruption frequency prediction model, we compare the observed number of disruptions (based on the empirical data set) with predicted numbers based on the MLP classifier. Predicted values are obtained from the 20% testing sample for each of the five folds in the fivefold cross validation applied to a 12-month data set, hence together providing predictions for one complete year. In Fig. 6, a comparison is shown between the predicted and observed disruption frequency for each disruption category, aggregated over stations and time periods for a complete year. There is a high correlation (> 0.99) between our predicted numbers and observed values. Especially predictions of exposure to door malfunctioning, brake malfunctioning, station overruns and infrastructure related disruptions are highly accurate. Notwithstanding, it can be noted that our prediction model tends to underestimate disruption exposure somewhat. On average the expected number of disruptions is underestimated by 5% using our model, indicating there is still potential for further model improvement. As our purpose is to ultimately predict disruption frequencies per station and time period, Fig. 7 provides validation results at a more disaggregate level. The absolute deviation between the yearly observed and predicted number of disruptions per station s and time period t is computed, for each disruption type separately. In Fig. 7, the percentage of cases is shown for which the absolute

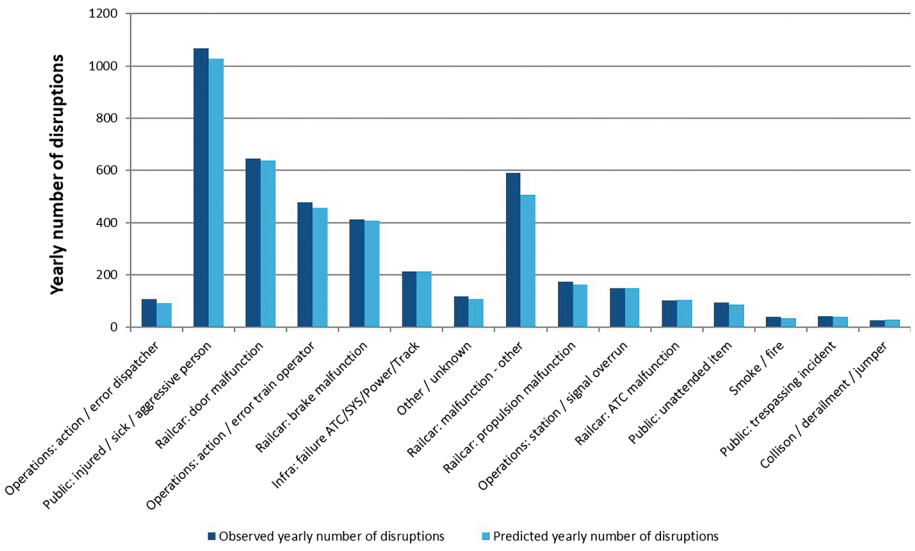

Fig. 6 Validation MLP classifier

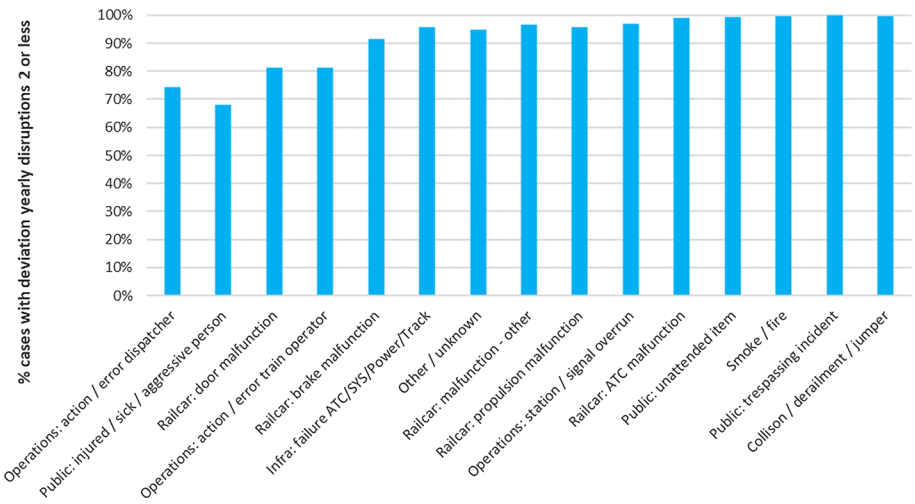

Fig. 7 Validation MLP classifier per station and time period

deviation between predicted and observed number of disruptions per station and time does not exceed 2 disruptions per year. For 10 of the 15 disruption types, 95% or more of the cases satisfy this condition. For 13 out of 15 disruption types, at least 80% of the cases satisfy this condition (with no value lower than 68% for any disruption type), which indicates that our model is also able to perform reasonably accurate predictions on a disaggregate level.

To validate our disruption impact prediction model, we compare the empirical passenger delay hours with predicted passenger delay hours using our Random Forest regression model. Predicted values are based on the 20% testing sample from each of the five folds used in the fivefold cross validation. This comparison is shown in Fig. 8, where the observed and predicted disruption impacts are aggregated over all stations and all disruptions per month. We can conclude there is a high correlation (>0.99) between predicted

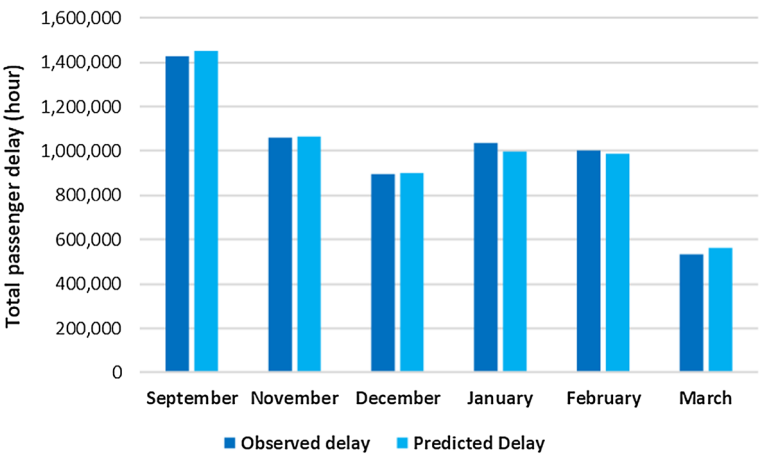

Fig. 8 Validation Random Forest regressor

and empirical delay hours. Per month, the predicted passenger delay deviates on average 0.6% from the empirical delay hours, with the maximum deviation per month being equal to 5.7% (March). The lowest deviation is observed for November, where the total predicted passenger delay deviates 0.5% from the total observed passenger delay. In Fig. 9, we further validate our model at a more disaggregate level, by comparing empirical and predicted disruption impacts for each individual disruption, station and time period in the dataset. It can be seen that passenger delay predictions deviate somewhat more from observed delays when assessed per individual incident. One probable cause for these deviations is the overlap between the impacts of some disruptions occurring close to each other in the dataset, making it more difficult to attribute delay impacts to individual disruptions. However, even for individual cases this deviation does not exceed 20% for 65–70% of the cases, whilst for 80% of the cases the deviations remain within a 30% range. These results give confidence that our proposed model is able to predict disruptions and passenger delay impacts reasonably well, although some bandwidth around predicted values may be incorporated in future developments. Hence, we can apply our models to predict the passenger delay impacts for each station, disruption type and time period. As empirical data about disruption frequency and impact is typically not available for all possible combinations of disruption type, station and time period, our models provide predictions for instances for which no or insufficient historical empirical data is available.

Prediction results

In Figs. 10 and 11, the feature importance is shown for the disruption exposure prediction model and disruption impact prediction model, respectively. As we predict probabilities for 16 different classes in our disruption exposure model, the feature importance is shown for each class separately. It can be seen that some features are particularly important in predicting the class *no disruption*, whereas other features are important in predicting one or more

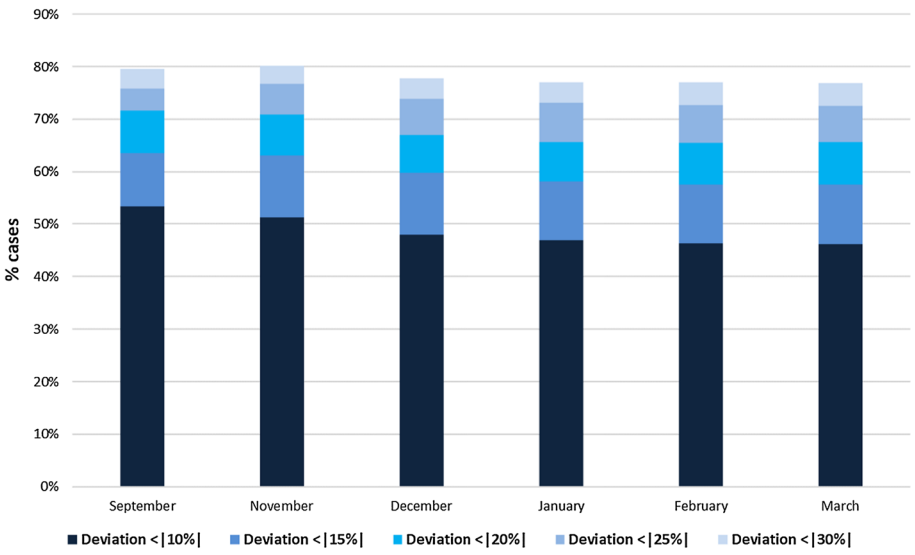

Fig. 9 Validation Random Forest regressor per disruption, station and time period

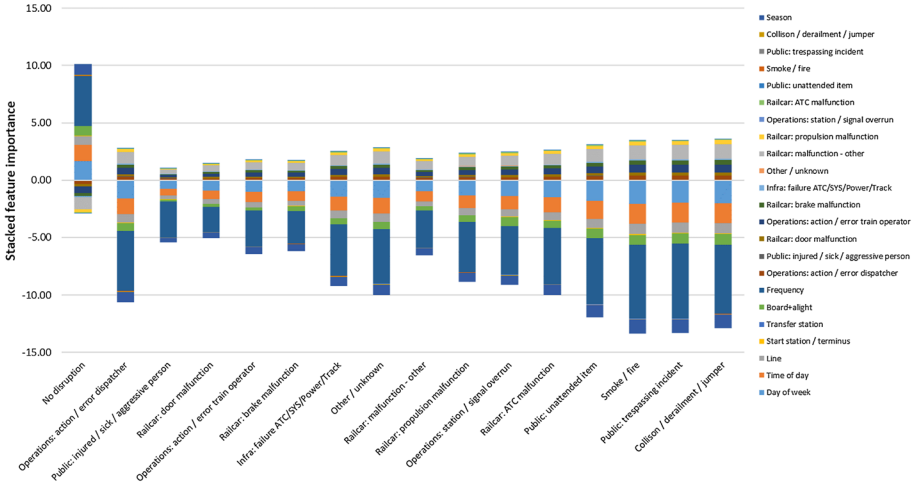

Fig. 10 Feature importance disruption exposure prediction model per class

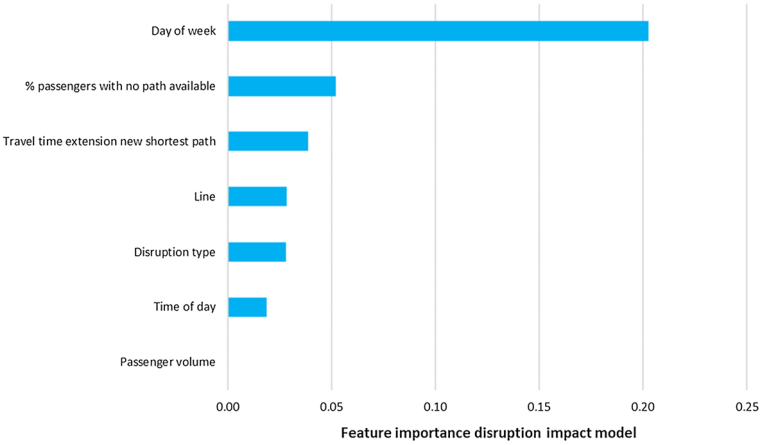

Fig. 11 Feature importance disruption impact prediction model

disruption classes. This entails that the feature importance might be positive for one class, whilst being negative for another class. Notwithstanding, features with negative importance for all classes are removed from the model. Figure 10 shows that *train frequency*, *day of the week* and *time of the day* are overall the three most important predictors for disruption exposure. For the disruption impact prediction model, Fig. 11 shows that *day of the week* (weekday/weekend day) is the most important feature in the Random Forest model. Additionally, the *percentage of the affected demand for which no path in the considered PT network remains available in case of a disruption* is also an important feature, followed by the *travel time extension of the new shortest path*. Interestingly, *passenger volume* has only a very minor contribution in predicting delay impacts in our model. Nevertheless, given the importance of the feature *percentage demand with no path available*, knowledge of passenger volumes is still necessary. As station-specific predictors are among the most important

predictors in both models, these results indicate the relevance of predicting disruptions and their passenger delay impacts for individual stations.

When combining the disruption exposure and impact prediction models, these models predict a yearly passenger delay of 5.9 million hours for the total metro network. This value is the sum of the expected station criticality of all metro stations in the network. Table 4 provides an overview of the 10 most and least critical stations with their expected contribution to the yearly passenger delay hours. For the most critical station *Gallery Place* the criticality equals almost 77,000 delay hours, whilst for the least critical station *Stadium-Armory* this value equals almost 43,000 delay hours per year. The 10 most critical stations are all located in the centre of the PT network, where train frequencies and passenger demand are highest. Five of all eight transfer stations in the network are positioned in this top-10 as well. The 10 least critical stations are all located on the eastern branch of the Orange or the Blue/Silver line, and on the western branch of the Silver line (see Fig. 2). None of these stops are start/terminal or transfer locations. Despite the limited number of route alternatives available to passengers when a disruption would occur at a station on one of these branches, the criticality of these stations is relatively low. Stations in the centre of the network are more often exposed to disruptions, and more passengers are affected once a disruption occurs. For stations in the centre section of our case study network, this suggests that the benefit of the availability of multiple route alternatives does not outweigh the costs, namely the more frequent disruption exposure and the larger passenger demand affected by these disruptions.

Clustering results

The station clustering result based on predicted criticality is shown in the dendrogram in Fig. 12. The cophenetic correlation coefficient equals 0.70, which can be considered reasonable. From the dendrogram can be seen that the 95 metro stations of the Washington metro network are grouped into five different clusters. Figure 13 shows for each of these clusters the expected yearly number of disruptions per station (left), the average (unweighted) disruption impact (centre), and expected yearly passenger delay per station (right). For stations in cluster 2 both the disruption exposure and impact are highest, resulting in the highest criticality. For this cluster, particular the average disruption impact is

Table 4 Station criticality ranking

Rank	Station (lines)	Criticality (pass-hours per year)	Rank	Station (lines)	Criticality (pass-hours per year)
1	Gallery Place (R)	76,594	86	Landover (O)	52,188
2	Metro Center (R)	74,384	87	Deanwood (O)	49,569
3	Gallery Place (YG)	72,653	88	Benning Road (SB)	49,438
4	Union Station (R)	72,439	89	Greensboro (S)	48,952
5	Metro Center (SOB)	71,926	90	Potomac Ave (SOB)	48,300
6	L'Enfant Plaza (YG)	70,314	91	Spring Hill (S)	48,095
7	Judiciary Square (R)	70,284	92	Cheverly (O)	45,696
8	Farragut West (SOB)	69,750	93	Capitol Heights (SB)	45,552
9	Columbia Heights (YG)	69,683	94	Morgan Boulevard (SB)	45,224
10	NoMa-Gallaudet U (R)	69,426	95	Stadium-Armory (SOB)	42,651

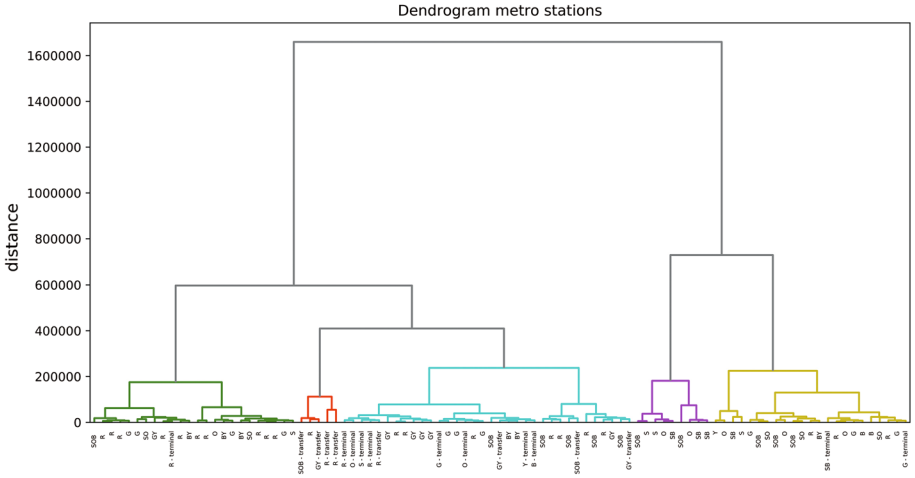

Fig. 12 Dendrogram with resulting clustering of metro stations. The labels on the x-axis show all 95 stations, with letters referring to the metro lines serving each station. Special station categories, such as transfer stations or terminals, are explicitly indicated. The distance on the y-axis shows the Euclidean distance between stations in terms of station criticality

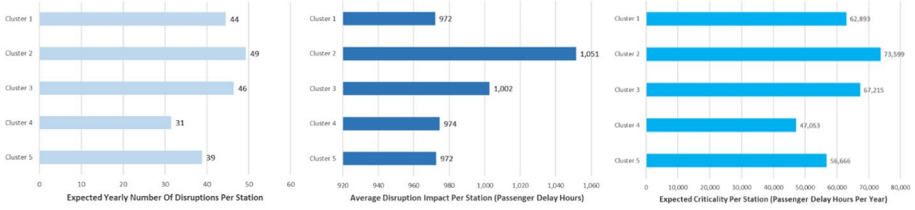

Fig. 13 Predicted average exposure (left), impact (centre) and criticality (right) per station in cluster

high compared to stations from other clusters. Stations from cluster 3 also have a relatively high disruption exposure, but the impact is smaller than for cluster 2. As a result, expected criticality for stations of cluster 3 is also lower than for cluster 2. The expected disruption impacts for stations from clusters 1, 4 and 5 are similar: these clusters differ primarily in their disruption exposure. Due to the relatively high disruption exposure in cluster 1 compared to clusters 4 and 5, the criticality of cluster 1 is also highest from these three clusters. Stations from cluster 4 are characterised by the lowest weakness, therefore resulting in lowest station criticality from all clusters.

In Fig. 14, we visualise the station ranking and clustering spatially. Each number in the figure shows the ranking of each of the 95 stations in terms of station criticality (see also Table 4), whereas the colour indicates the cluster to which each station belongs. Cluster 2 (Fig. 12, red) consists of five stations: four transfer stations and the main train station Union Station. These stations are most critical, as exposure and impact are highest. The high passenger delays characterising this cluster can be explained by the relatively high number of passenger-related and railcar-related disruptions, due to the high train frequency and passenger volumes in this central part of the metro network. These high passenger volumes also result in the highest disruption impacts. Cluster 3 (Fig. 12, blue) is the largest cluster,

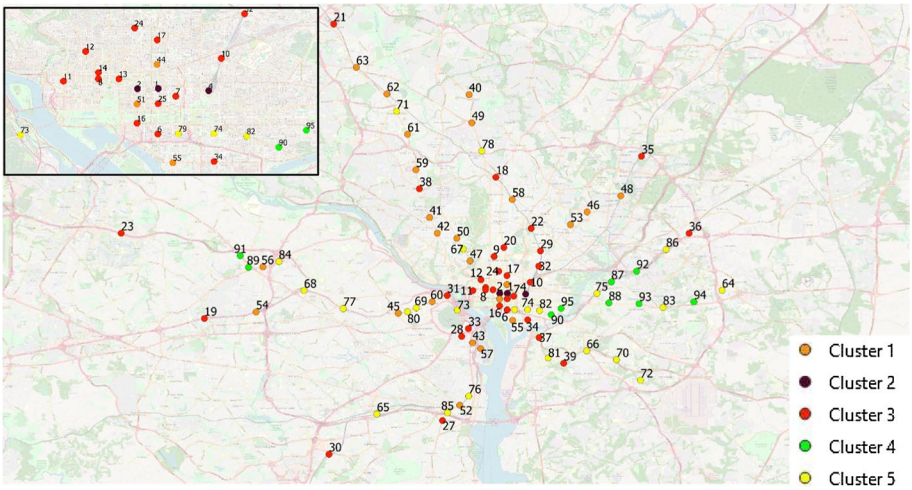

Fig. 14 Station criticality ranking and clustering. Numbers refer to station criticality ranking 1-95; colours refer to the five clustering categories, ranging from green (cluster with lowest criticality) via yellow, orange and red to dark purple (cluster with highest criticality). The inset upper-left zooms into the centre part of the network

containing 34 stations. The criticality of these stations is second-highest, after the stations from cluster 2. This cluster consists of all remaining transfer stations and the majority of the start/terminal stations. Besides, most other stations located in the centre area of the PT network are part of this cluster. All these stations are relatively heavily exposed to disruptions. For stations in the centre part of the network, this is mainly caused by the high train frequencies and passenger volumes. For the start/terminal stations, an explanation is that several disruptions often arise at the first station of the line: a railcar malfunctioning when testing the train, a late or sick driver not arriving on time, or a late movement of the train from the yard to the first station. As confirmed from the empirical analysis (Fig. 4), start/terminal stations are more frequently exposed to disruptions than their surrounding stations. As passenger demand at the stations of cluster 3 is lower than for the five busy stations of cluster 2, the expected disruption impact is lower as well. Cluster 4 (Fig. 12, purple) contains the 9 least critical stations. As shown in Table 4, these stations are located at the end of the western branch of the Silver line, and at the end of the eastern branches of the Orange and Blue/Silver lines. These stations have the lowest disruption exposure from all stations for all disruption types, particularly in relation to passenger-related disruptions. One can argue that the relatively low headways combined with relatively low passenger volumes at the end of these lines result in lower passenger impacts, despite the number of available route alternatives in the network also being smaller here. Location-specific characteristics might also play a role here. The stations of cluster 1 (Fig. 12, green) and cluster 5 (Fig. 12, gold) are mainly located between the busiest centre section of the network and the outer branches of the lines. The 24 stations of cluster 1 are primarily stations at the northern part of the network (Red line, trunk section of the Yellow/Green lines, northern branch of the Green line). The 23 stations of cluster 5 are mostly located at the southern part of the network (trunk section of the Silver/Orange/Blue lines, trunk section of Blue/Yellow lines, southern branch of the Green line). In terms of exposure, these clusters can be positioned between clusters 2/3 on the one hand, and cluster 4 on the other hand.

Especially stations from cluster 1 are relatively often exposed to disruptions: this might be caused by the relatively high frequency on the Red line, or by differences in operating stock types. Disruptions at stations of clusters 1 and 5 affect more passengers compared to cluster 4, but some stations offer more route alternatives to these affected passengers. These positive and negative effects seem to cancel out each other on average, as the average disruption impact is similar to stations from cluster 4.

Conclusions

In this study we propose a generic approach to predict how often different disruption types occur at different stations of a PT network, and to predict the impact related to these disruptions as measured in terms of passenger delays. The contribution of our research lies in the development of supervised learning models to predict disruptions and their passenger impacts for each individual station, disruption type and time period, as sufficient empirical disruption observations will not always be available for each location and time period. Besides, our models can predict disruption impacts for all stations and time periods for a medium-sized PT network (consisting of 95 metro stations) within 10 min. Hence, our method provides an alternative for existing, computationally more expensive methods to predict passenger delays for a complete PT network. Applied to the Washington metro network, our models predict a yearly passenger delay of 5.9 million hours for the total metro network. Five different types of station are distinguished by clustering stations according to their expected criticality. Stations with high train frequencies and high passenger volumes located at central trunk sections of the network show to be most critical, together with start/terminal and transfer stations. Intermediate stations located at branches of a line are least critical. The lower train frequencies and passenger volumes result in lower disruption exposure and impact, despite less route alternatives typically being available for these passengers when a disruption occurs.

Our study results provide PT authorities and service providers insights into the frequency, location and passenger impact of different disruptions. It provides an overview of the stations which contribute most to the vulnerability of the total PT network. Categorising stations based on their disruption characteristics shows the different station types which can be distinguished based on their contribution to network vulnerability. This supports PT agencies in prioritising what type of disruptions at what location to focus on, to potentially achieve the largest improvements in network robustness. Ranking all stations according to their criticality directly supports decision-makers to target robustness measures at these stations which need it most. The explicit distinction between disruption exposure and impact helps determining what type of measure would be most suitable for each (type of) station. Our method can also be used to quantify the robustness benefits of new infrastructure, such as a new rail link. The model trained for the current PT network can be used to predict the new station criticality in the event of a network adjustment, by updating network-related predictors. This results in a fast and complete quantification of robustness benefits, which can be incorporated in appraisal studies.

We formulate four recommendations for future research. First, we recommend further testing of our model sensitivity in relation to missing disruption duration information. We recommend to apply this method to other case study networks, where the disruption duration—possibly including recovery time—is provided in the disruption log data, so that the sensitivity of the model performance can be investigated. This might enable a further

improvement of the R^2 value of the disruption impact model, which can currently be considered reasonably high. Second, application of our method to a link based or joint node and link based vulnerability analysis is recommended. As disruptions in the data set provided to us were allocated to stations, we applied a node based vulnerability analysis. Our methodology is however directly applicable for link based analyses using the same predictors. Testing the sensitivity of our model outcomes to this is therefore recommended. Third, although our model allows for a reasonably accurate prediction of disruption impacts, our model slightly underestimates disruption predictions by 5% on average. Future research is therefore recommended to further improve of the accuracy of this prediction model. Fourth, we also recommend incorporating the availability of other modes in the assessment of the number of paths remaining available for passengers, as well as for the indication of passengers' travel time extension, as used as predictors in our disruption impact prediction model. As our model only considers metro lines, robustness resulting from the availability of alternative modes of transport in the network is potentially somewhat underestimated. In our study we compare the difference between realised and scheduled journey time from the first metro station at which a passenger tapped in, to the final metro station where one has tapped out. Therefore, we do account for potential denied boarding in our delay calculation. This however implies that disruption impacts on *perceived* journey times (e.g. due to increased crowding levels) are not captured in our model. In addition, when passengers would change their boarding or alighting station in response to a disruption, this could result in longer access or egress times to and from the metro network. This delay impact outside the considered metro network itself is not captured in our current model, but is recommended to consider in future research.

Acknowledgements This research was performed as part of the TRANS-FORM (Smart transfers through unravelling urban form and travel flow dynamics) project funded by NWO Grant agreement 438.15.404/298 as part of JPI Urban Europe ERA-NET CoFound Smart Cities and Communities initiative. The authors thank WMATA, and Jordan Holt in particular, for the valuable cooperation and data provision.

Authors' contribution Both authors contributed to the research design and methodology, to the analysis of the case study results, and to the writing of the paper.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest. Parts of this paper are based on an earlier conference paper by Yap and Cats (2019). In this conference paper, we propose a method to predict disruption frequencies in metro networks. In the current paper, we have improved our models to predict disruption frequencies. Besides, we extended our modelling framework by developing new methods to predict passenger delay impacts of disruptions, and rank and cluster stations based on both disruption frequency and impact characteristics.

References

Bell, M.G.H.: The use of game theory to measure the vulnerability of stochastic networks. IEEE Trans. Reliab. **52**, 63–68 (2003)

Byrd, R., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. J. Sci. Comput. **16**, 1190–1208 (1995)

Cats, O., Jenelius, E.: Dynamic vulnerability analysis of public transport networks: mitigation effects of real-time information. Netw. Sp. Econ. **14**, 435–463 (2014)

Cats, O., Jenelius, E.: Planning for the unexpected: the value of reserve capacity for public transport network robustness. Transp. Res. Part A **81**, 47–61 (2015)

Cats, O., Jenelius, E.: Beyond a complete failure: the impact of partial capacity degradation on public transport network vulnerability. Transp. A **6**, 77–96 (2018)

Cats, O., Yap, M.D., Van Oort, N.: Exposing the role of exposure: public transport network risk analysis. Transp. Res. Part A **88**, 1–14 (2016)

Corman, F., D'Ariano, A., Hansen, I.A.: Evaluating disturbance robustness of railway schedules. J. Intell. Transp. Syst. **18**, 106–120 (2014)

Derrible, S., Kennedy, C.: The complexity and robustness of metro networks. Phys. A **389**, 3678–3691 (2010)

Dinh, T.N., Thai, M.T.: Network under joint node and link attacks: vulnerability assessment methods and analysis. IEEE/ACM Trans. Netw. **23**, 1001–1011 (2014)

Geurs, K.T., La Paix, L., Van Weperen, S.: A multi-modal network approach to model public transport accessibility impacts of bicycle-train integration policies. Eur. Transp. Res. Rev. **8**, 1–15 (2016)

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.M., Attanucci, J.P.: Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp. Res. Rec. **2343**, 17–24 (2013)

Hörcher, D., Graham, D.J., Anderson, R.J.: Crowding cost estimation with large scale smart card and vehicle location data. Transp. Res. Part B **95**, 105–125 (2017)

Jenelius, E.: Incorporating dynamics and information in a consequence model for road network vulnerability analysis. In: Proceedings of Third International Symposium on Transport Network Reliability (INSTR), The Hague, The Netherlands (2007)

Kingma, D., Ba, J.L.: Adam: A method for stochastic optimization. In: Proceedings of the ICLR Conference, San Diego (2015)

Knoop, V.L., Hoogendoorn, S.P., Van Zuylen, H.J.: The influence of spillback modelling when assessing consequences of blockings in a road network. Eur. J. Transp. Infrastruct. Res. **8**, 287–300 (2008)

Knoop, V.L., Snelder, M., Van Zuylen, H.J., Hoogendoorn, S.P.: Link-level vulnerability indicators for real-world networks. Transp. Res. Part A **46**, 843–854 (2012)

Lord, D., Washington, S.P., Ivan, J.N.: Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. **37**, 35–46 (2005)

Malandri, C., Fonzone, A., Cats, O.: Recovery time and propagation effects of passenger transport disruptions. Phys. A **505**, 7–17 (2018)

Munizaga, M.A., Palma, C.: Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C **24**, 9–18 (2012)

Murray, A.T., Matisziw, T.C., Grubesic, T.H.: A methodological overview of network vulnerability analysis. Growth Change A J. Urban Reg. Policy **39**, 573–592 (2008)

Oliveira, E.L., Silva Portugal, L., Porto Junior, W.: Indicators of reliability and robustness: similarities and differences in ranking links of a complex road system. Transp. Res. Part A **88**, 195–208 (2016)

Pedregoas, F., Varoquaux, G., Gramfort, A., Michel, V.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

Rodriguez-Nunez, E., Garcia-Palomares, J.C.: Measuring the vulnerability of public transport networks. J. Transp. Geogr. **35**, 50–63 (2014)

Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**, 83–112 (2017)

Scott, D.M., Novak, D.C., Aultman-Hall, L., Guo, F.: Network Robustness Index: a new method for identifying critical links and evaluating the performance of transportation networks. J. Transp. Geogr. **14**, 215–227 (2006)

Stone, M., Broughton, J.: Getting off your bike: cycling accidents in Great Britain in 1990–1999. Accid. Anal. Prevent. **35**, 549–556 (2003)

Sullivan, J.L., Novak, D.C., Aultman-Hall, L., Scott, D.M.: Identifying critical road segments and measuring system-wide robustness in transportation networks with isolating links: a link-based capacity-reduction approach. Transp. Res. Part A **44**, 323–336 (2010)

Tampère, C.M.J., Stada, J., Immers, B., Peetermans, E., Organe, K.: Methodology for identifying vulnerable sections in a national road network. Transp. Res. Rec. **2012**, 1–10 (2007)
Tirachini, A., Hurtubia, R., Dekker, T., Daziano, R.A.: Estimation of crowding discomfort in public transport: results from Santiago de Chile. Transp. Res. Part A **103**, 311–326 (2017)
Tonnelier, E., Baskiotis, N., Guigue, V., Gallinari, P.: Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework. Neurocomputing **298**, 109–121 (2018)
Yap, M.D., Cats, O.: Analysis and prediction of disruptions in metro networks. In: Proceedings of the 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Krakow (2019)
Yap, M.D., Cats, O., Van Oort, N., Hoogendoorn, S.P.: A robust transfer inference algorithm for public transport journeys during disruptions. Transp. Res. Proc. **27**, 1042–1049 (2017)
Yap, M.D., Cats, O., Van Arem, B.: Crowding valuation in urban tram and bus transportation based on smart card data. Transp. A (2018a). https://doi.org/10.1080/23249935.2018.1537319
Yap, M.D., Van Oort, N., Van Nes, R., Van Arem, B.: Identification and quantification of link vulnerability in multi-level public transport networks: a passenger perspective. Transportation **45**, 1161–1180 (2018b)
Zhang, X., Guo, C., Wang, L.: Using game theory to reveal vulnerability for complex networks. In: Proceedings of the 10th IEEE International Conference on Computer and Information Technology, Bradford (2010)
Zou, X., Yue, W.L.: A Bayesian network approach to causation analysis of road accidents using Netica. J. Adv. Transp. (2017). https://doi.org/10.1155/2017/2525481