Transportation Research Part C 144 (2022) 103880

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Detecting metro service disruptions via large-scale vehicle location data

Nan Zhang [a], Daniel J. Graham [a,*], Prateek Bansal [b], Daniel Hörcher [a]

[a] Transport Strategy Centre, Department of Civil and Environmental Engineering, Imperial College London, London, UK
[b] Department of Civil and Environmental Engineering, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Metro system
Disruption detection
Gaussian mixture model
Secondary disruption

ABSTRACT

Urban metro systems are often affected by disruptions such as infrastructure malfunctions, rolling stock breakdowns and accidents. The crucial prerequisite of any disruption analytics is to have accurate information about the location, occurrence time, duration and propagation of disruptions. To pursue this goal, we detect the abnormal deviations in trains' headway relative to their regular services by using Gaussian mixture models. Our method is a unique contribution in the sense that it proposes a novel, probabilistic, unsupervised clustering framework and it can effectively detect any type of service interruptions, including minor delays of just a few minutes. In contrast to traditional manual inspections and other detection methods based on social media data or smart card data, which suffer from human errors, limited monitoring coverage, and potential bias, our approach uses information on train trajectories derived from automated vehicle location (train movement) data. As an important research output, this paper delivers innovative analyses of the propagation progress of disruptions along metro lines, which enables us to distinguish primary and secondary disruptions as well as effective recovery interventions performed by operators.

1. Introduction

With high-frequency services and large capacity, metros (also known as subways or rapid transit) play a vital role in transporting the urban population. However, large-scale urban metro systems are vulnerable to disruptions, which cause passenger delays, crowding concerns and can negatively affect passenger satisfaction with metro operations. In this study, we specifically focus on service disruptions[1] which are defined as events that interrupt normal train operations for a specific duration.[2] Each event is indicated by an abnormal deviation (or delay) in trains' headway relative to their scheduled services at a metro platform. These disruptions are often caused by unpredicted infrastructure malfunctions (e.g., signal failures and track blockages), rolling stock breakdowns and accidents, planned maintenance work, and temporal dispatching adjustments (Jespersen-Groth et al., 2009). To quantify and improve

* Corresponding author.
E-mail addresses: d.j.graham@imperial.ac.uk (D.J. Graham), prateekb@nus.edu.sg (P. Bansal), d.horcher@imperial.ac.uk (D. Hörcher).

[1] To distinguish from the broader term "incidents" or "anomalies", service disruptions do not include events that do not affect train services. For example, the escalator failure or corridor congestion in metro stations are not service disruptions.

[2] In general, service disruptions may last from a few minutes to several hours or days, depending on the severity of the disturbance to service provision. The disruption duration is defined as the time beyond the regular headway deviations that happen due to minor fluctuations in daily train operations.

https://doi.org/10.1016/j.trc.2022.103880
Received 20 December 2021; Received in revised form 17 May 2022; Accepted 2 September 2022
Available online 27 September 2022
0968-090X/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

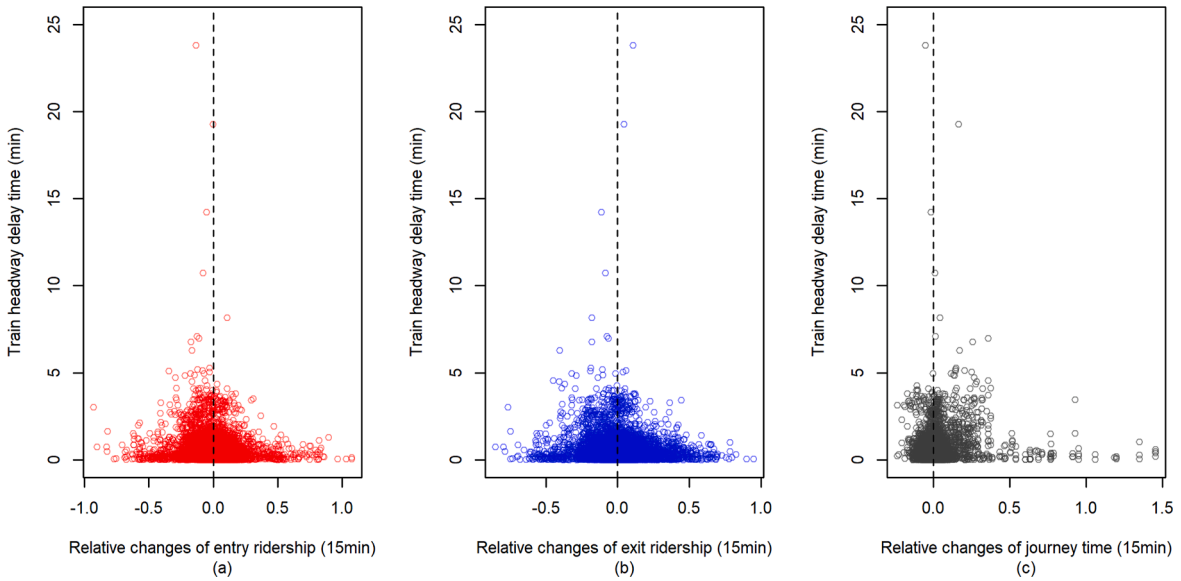

Fig. 1. The relationship of abnormal travel patterns and service delays for the example station.

a system's reliability, metro operators need detailed information about service disruptions, including the locations, occurrence time, durations and network propagation. As an essential input to disruption management, this information can help operators prepare effective recovery plans and guide future maintenance. On the other hand, for metro users, real-time service disruption information is also an integral component of the advanced passenger information system. With the latest updates regarding the occurred disruptions, delayed status and estimated recovery time, passengers are better able to reschedule their trips under unexpected situations. Thus, this research develops a data-driven method to accurately detect service disruptions and their propagation along the network.

Traditionally, to detect disruptions metro operators rely on reports from manual inspections and complaints from passengers. Such detection results usually suffer from human errors and are restricted to a limited monitoring range in both space and time due to resource constraints (Ji et al., 2018). Therefore, recent studies have used two new data sources to identify disruptions. First, Ji et al. (2018) and Zulfiqar et al. (2020) leverage social media data such as tweets with the keywords of metro lines, stations and common complaint vocabulary to predict disruptions. Although social media data can capture a significant amount of passengers' feedback and cleverly monitor metro disruptions in spatiotemporal dimensions, human errors cannot be circumvented in this approach. Second, a few studies have mined automated fare collection or smart card data (SCD) to capture abnormal passenger behaviour and assume that uncommon travel patterns such as anomalous change of station ridership and extra journey time are good indicators of incident occurrence (Sun et al., 2016; Tonnelier et al., 2018; Briand et al., 2019; Jasperse, 2020). However, such indicators may not be ideal for detecting service disruptions because, instead of train interruptions, other factors can also significantly affect passenger behaviour and corresponding demand measures. For instance, adverse weather conditions and external mega-events (e.g., concerts or sports matches) may cause demand fluctuations.

To validate the above hypothesis (service disruption is related to unusual travel patterns) with data, in Fig. 1 we plot the relationship between the abnormal changes in passenger demand/journey times[3] and service delays[4] for a busy station of a major metro system in Asia. As shown in the three plots, there were several significant delays of more than 10 min when demand and journey time remained largely unchanged. Conversely, when significant changes in ridership and journey time were observed, headway delays tended to be less than 5 min. Thus, the relationship between service disruptions and travel patterns cannot be supported by the data.[5] To strengthen this view, statistical hypotheses (H_0 : the correlation coefficient $= 0$, $\alpha = 0.05$) have been tested for the correlations between headway delays (over 3 min) and the relative changes in travel patterns. The p-values for entry/exit ridership and average journey time are all greater than 0.05. Overall, no significant correlations have been found. Therefore, although SCD are useful in detecting demand-related incidents in the system, an aberrant change in demand or journey time is not necessarily a sign of service disruption. There is a need to explore the potential of other emerging datasets and methods to detect service disruptions. We focus on

[3] Relative change in demand = (the entry or exit demand for a 15 min interval − the 15 min average demand)/(the average demand of that 15 min across all observed weekdays).Relative change in journey time = (the average journey time for passengers who start traveling in a 15min interval − the 15min mean average journey time)/(the mean average journey time of that 15min across all observed weekdays).For the example station, Fig. 1 includes all data points observed on 54 weekdays during the study period. The 15-minute interval for demand calculation begins when the corresponding headway delay is first observed.

[4] The service delay time is determined as the (positive) difference between the observed headway and the scheduled headway.

[5] If the given hypothesis is supported, we would have observed increase in headway deviations due to increase in demand/journey time.

service (train) delays because information on the propagation of such delays is more useful in informing the corresponding control measure as it can be directly integrated with schedule optimisation models. Historical delay data have been applied to optimise timetables, disturbance-recovery strategies, and energy consumption of metro systems (Yang et al., 2019; Li et al., 2020).

This research proposes a novel approach to detect service disruptions using large-scale vehicle location data. The proposed method is free of human errors and would enable an analyst to investigate service disruptions across spatial and temporal dimensions. The deviation in headway relative to the scheduled headway is used as the indicator of disruption occurrence. The proposed method involves two steps – (i) split the day into 30-minute intervals and detect whether the platform is disrupted during a specific interval, and (ii) identify the propagation of disruption across the metro line over time.

First, we apply a Gaussian mixture model (GMM) on the headway deviations to identify a cluster of abnormal headways. The probability of a station being associated with the cluster of abnormal headways during a specific interval is the by-product of the estimation. To convert the disruption probabilities into final detection decisions, the optimal probability threshold (i.e., minimum probability of observation to fall in the abnormal cluster to be called disrupted) is also learned from a simulation-based method. The proposed probabilistic unsupervised learning framework thus obviates the need to subjectively define anomaly thresholds.

Second, by merging the detection output of the first step with the train trajectory data at the line level, propagation of the disruption across the connected stations is identified. In this way, we can identify the station with the origin of disruption (i.e., primary disruption) and the extent of the spill-over interruption on downstream/upstream platforms (i.e., secondary disruption). Our approach involves a smart screening algorithm, followed by visualisation of disruptions on the space–time diagram of train movement. These diagrams also reveal the effective recovery interventions performed by metro operators, such as dispatching adjustments or rescheduling. In the urban metro sector, this is the first time that secondary disruptions and recovery interventions are identified comprehensively in raw data with empirical methods, which offers an alternative to manual detection and simulation-based approaches The former reflects the impact of the primary disruption on the service provision of downstream stations, which is essential to comprehensively evaluate line-level reliability. The latter reflects the ability of operators to restore normal services under disruptions. Quantifying recoverability plays a key role in assessing metro resilience.

We use a major metro system in Asia as a case study and apply the proposed method to detect service disruptions on a densely used line. Compared to manual incident logs, we have detected all disruptions (over 5 min) and 96 % of minor incidents (between 2 and 5 min). In terms of the validation via simulated detections, across all stations of the selected metro line under both minor and mixed disruption scenarios, the average detection accuracy is above 0.99. Specifically, the average precision is nearly uncompromised, and the average recall rate is over 0.9. The detection results contain detailed information of historical disruptions, including their occurrence time and location, lasting duration as well as the propagation of service delays along metro lines. Such information is the foundation for further research on disruption impacts and management, with which operators can optimise recovery strategies and dynamic scheduling. Accurate service delay information also improves the evaluation of service reliability.

The rest of the paper is organised as follows. Section 2 reviews the literature on disruption detection in transport systems, with a focus on urban metro systems. Section 3 presents our probabilistic framework to detect train service interruptions. This section discusses GMM, followed by the screening algorithm and space–time diagram approach to identify secondary disruptions and effective recovery interventions. We carry out an empirical analysis to detect disruptions in Section 4, and results are detailed in Section 5. Finally, conclusions and potential avenues for future research are discussed in Section 6.

2. Literature review

In the road traffic sector, anomaly detection has been widely analysed. The concept of anomaly includes abnormal traffic conditions such as accidents, congestions (disruption of road services) or specific road-related events. The detection methods have evolved from constant human observance through CCTV monitors to automatic detection based on sensors and algorithms (Mahmassani et al., 1999). The exploited sensor data include inductive loop detectors data (Rossi et al., 2015; Zhu et al., 2018), social media data (Gu et al., 2016; Zhang et al., 2018), GPS trajectory data (D'Andrea and Marcelloni, 2017; Yu et al., 2020; Zhang et al., 2021), camera data (Cano et al., 2009; Sodemann et al., 2012; Riveiro et al., 2017; Santhosh et al., 2020) and mobile phone data (Steenbruggen et al., 2016; Bolla and Davoli, 2000), among others. The above practices have shown that the detection of traffic anomalies is a data-specific task. Specifically, the choice of detection method depends on the structure and characteristics of the data used.

In railway systems, the concept of train delay has been widely used as an indicator of service disruptions. There are three common methods to determine train delays and their prorogation: the analytical approach, micro-simulation approaches and statistical analyses based on empirical data (Mattsson, 2007). The analytical models generally rely on queueing theory and require simplifying assumptions due to limited information about the railway system (Milinković et al., 2013). Under a set of timetable-structure assumptions, Harrod et al. (2019) proposed a closed form formulation of delay propagation to provide analytical estimate of aggregate service delay. Goverde (2010) converted a scheduled railway as a discrete-event dynamic system. The use of max-plus algebra does allow analytical modelling of delays over a periodic timetable. Simulation models aim to imitate delayed situations of a railway system with more granularity. With data regarding the infrastructure design, the timetable and the performance of the trains, interactions between the trains and the infrastructure can be quantified in simulated scenarios (Hansen, 2008). Carey and Kwieciński (1994) developed stochastic approximations to measure delay propagation, which were tested and calibrated via detailed stochastic simulation. The simulation framework of train delay research is also critical in the optimisation of dispatching and capacity utilization (Yuan and Hansen, 2007; D'Ariano and Pranzo,2009; Dollevoet et al., 2014; Corman et al., 2017). As for empirical analyses, train delays can be measured by filtering track occupation and release records and comparing with scheduled arrival and departure times of a dispatching plan at reference points (Hansen et al., 2010). Statistical models can be used to identify the source or type of delays,

estimate the probability distribution of primary and knock-on train delays, and analyse the delay propagation (Yuan, 2006). For instance, regression models are often used to find the significant factors of arrival, departure, and dwell time delays (Dingler et al., 2010). Milinković et al. (2013) proposed a fuzzy Petri net model to estimate train delays with large external disturbances in the system of Serbian Railways. Büker and Seybold (2012) presented a formalisation of delay propagation via stochastic modelling of activity graphs. Meester and Muns (2007) argued that, based on phase-type distributions, it is possible to derive secondary delay distributions from primary delay distributions. Similarly, Yuan (2006) proposed a stochastic model to predict the distribution of departure delays according to the distribution of late arrival and dwell delays using Monte Carlo sampling method.

Although train delay is a viable indicator for disruption detection in mixed-use inter-urban railway systems, it does not provide the same information in high-frequency urban metro systems. In metros, maintaining a planned service frequency is the primary objective and train-specific operational timetables are difficult (and not necessary) to maintain. Thus, a deviation of the train's location from a pre-planned timetable may not indicate a metro disruption, while abnormal headways do. Recent studies have started to apply advanced machine learning methods to predict railway delays and their propagation, such as deep learning frameworks, neural networks and deep and shallow extreme learning machines (Yaghini et al., 2013; Oneto et al., 2018; Huang et al., 2020; Huang et al., 2021; Spanninger et al., 2022). Zhou et al. (2020) forecast the change of train delays and propagation by using the random vector functional-link networks with improved transfer learning and ensemble learning. These methods aim to estimate the expected primary or knock-on train delays based on the pre-defined timetable, train movement data, railway infrastructure and historical delays. Unfortunately, none of the advanced approaches can be directly applied to metro systems, as the ground truth of historical service disruptions and the primary/secondary categorisations are not available at the current stage.

In the urban rail sector, huge efforts have been made to automatically detect faults in track circuits. Using circuit sensor signals and video images, researchers have developed diagnostic algorithms based on neural network, deep learning and Bayesian network to automatically detect faults in railway track circuits (Chen et al., 2008; Zhao et al., 2012; De Bruin et al., 2017; Ashish et al., 2018; Welankiwar et al. 2018; Wei et al., 2019). Due to the adoption of different data sources, such methods are not transferrable to our context.

However, most urban metro systems still rely on manual incident detection methods, which are based on reports of manual inspections from metro operators and complaints from passengers (Ji et al., 2018). For example, when a disruption occurs in the London Underground, the staff involved are required to complete an Incident Reporting Form (IRF). After verification by an operational manager, the IRF is entered into the service data system called CuPID, which finally generates incident logs (London Datastore, 2018). These traditional detection methods suffer from human errors and manpower constraints, leading to missing observations due to limited monitoring range in space and time and incorrect records (Ji et al., 2018). Such shortcomings of traditional methods have encouraged researchers to explore automatic disruption detection methods in urban metro systems by leveraging the emergence of large-scale datasets.

The first type of new data source is social media. Metro-related social media data include reviews or comments made by passengers about metro services. For instance, Collins et al. (2013) applied Twitter data to evaluate the satisfaction level of riders with transit service. Osorio-Arjona et al. (2021) also used Twitter data to extract complaints or negative feelings related to Madrid Metro system. They discovered the spatial distribution of complaining users. Luo and He (2021) utilised data related to the Shenzhen metro system from a Chinese microblogging website to explore the points of interest of the transit users, such as crowding level, service reliability and waiting time. Ni et al. (2016) used Twitter data to predict passenger volume at a subway station in New York City. As for the application of social media data under incident occurrences, Ji et al. (2018) used Twitter data to detect service disruptions in the Washington Metro. The authors first filtered tweets with keywords of metro lines and stations during a given period. Subsequently, by mining common complaint vocabulary in these tweets (fail, disrupted, interrupted, and injury, among others), they predicted if there was a delay on a specific metro line. Similarly, Zulfiqar et al. (2020) used real-time Twitter data to develop an open-source system for the early detection of emergencies or criminal events within rail-based transit systems. By tracking the emerging information about each particular incident, they are able to track the chronological development of threatening events during the day. Compared with conventional incident logs data, the social media data can capture passengers' feedback and complaints promptly and cleverly monitor train services throughout the entire network. However, the detection of disruptions based on social media data cannot avoid the limitations of human inspections. For example, it might be the cases that not all disruptions are mentioned on social media, line/station information might be missing, and the posts may contain wrong or fake disruption information. Thus, the detection accuracy largely depends on the representativeness and quality of metro-related social media data.

The second type of new data source is automated fare collection or smart card data (SCD). SCD provides information about the origin and destination of passengers with timestamps, and thus, reveals the journey time and travel behaviour of passengers. Sun et al. (2016) used SCD to obtain passenger flows and regarded abnormal changes in passenger flow as a sign of disruption occurrence. They assumed that the passenger arrival rate during a specific period follows a normal distribution and estimated the distributional parameters via Bayesian inference. Subsequently, they considered all observations beyond three standard deviations of the mean passenger flow as the disruption indicator. In another study, Tonnelier et al. (2018) proposed four approaches for anomaly detection using SCD. The first three approaches inferred a daily temporal prototype (that is, a specific pattern of passenger behaviour depending on the day of the week) according to entry logs using three different methods: the average, the normalised average and a discrete probability density function obtained from the nonnegative matrix factorisation algorithm. Next, they obtained anomaly scores by determining the

Table 1
A comparison of recent research on metro disruption detection.

Research	Source data	Detection indicator	Detection method	Detection accuracy		Disruption propagation	Recovery intervention
				No human errors	True service disruption		
Sun et al., 2016	Smart card data	Boarding ridership	Three-standard-deviation rule with Bayesian inference	√			
Ji et al., 2018	Twitter data	Complain/delay vocabulary in tweets	Multitask supervised learning		√		
Tonnelier et al., 2018	Smart card data	Entry logs	Anomaly scores compared to baseline	√			
Briand et al., 2019	Smart card data	Passenger demand	Boxplot	√			
Jasperse, 2020	Smart card data	Passenger delay	Hierarchical clustering and probabilistic classification	√		*[1]	
Zulfiqar et al., 2020	Twitter data	Crime/emergency vocabulary in tweets	Based on keywords and dynamic query expansion				
Our approach	Vehicle location data	Service headway	Probabilistic Gaussian mixture model	√	√	√	√

[1] Jasperse (2020) analysed the spread of passenger delays, rather than the propagation of train service delays.

difference between the inferred prototypes of station flow and the real observations of a particular day. The fourth approach is a user-based model in which they computed the log probability distribution of the entry frequencies at different stations for each passenger. Abnormal entry behaviour was detected at the passenger level and aggregated in spatiotemporal dimensions. Briand et al. (2019) used the unexpected increases or decreases in passenger demand to detect atypical events in the French transit network of the city of Rennes. After clustering observations with similar ridership activities, they conducted outlier detection based on the boxplot method. Jasperse (2020) also used SCD, but he relied on abnormality in journey time rather than demand patterns to identify irregular metro operations. The author attempted to detect the propagation of passenger delays in both spatial (that is, through nearby stations) and temporal dimensions. However, this study focuses on passenger delays rather than train service delays. In metro systems, service delays would generally result in passenger delays. But on the contrary, service delays cannot be directly inferred from passenger delays, since passenger delays may originate from other factors, such as overcrowding. This distinction is crucial.

In summary, SCD-based methods might omit some service disruptions because abnormal passenger behaviour is not significantly correlated with the abnormal headways. Such weaker correlations can be attributed to the fact that anomalous travel patterns can be caused by many other factors apart from service disruptions: (i) inherent fluctuations in the passenger demand itself, (ii) weather conditions, (iii) mega-events near metro stations, and (iv) temporary demand control measures. Thus, SCD could be useful in detecting ridership related incidents, but the detection of service disruptions requires a new data source and method, unless the above causes of anomalous travel patterns can be fully controlled. Table 1 shows a comparison of recent research on metro disruption detection.

Previous studies on the propagation of urban metro delays are limited to simulation experiments. Malandri et al. (2018) used simulated disruption scenarios to measure the change of passenger volume over capacity ratio through the network in both space and time. Also based on simulations, Yap et al. (2022) tested different rescheduling strategies with dynamic passenger assignment to minimise the overall user delays caused by the hypothetical primary disruptions. Under the optimal strategy, they estimated the arrival delay of rescheduled trains and the propagation of disruption impacts on journey time and costs. However, we are not aware of any empirical study on the propagation of abnormal status of service provision.

To identify effective recovery interventions under disruptions, simulation-optimisation frameworks are commonly used. Tessitore et al. (2022) used a metro traffic simulator to evaluate the performance of several operational interventions for traffic disturbance recovery in metro systems. Gkiotsalitis and Cats (2020) proposed an exact model for timetable recovery and carried out multiple experiments to investigate how the upstream trips should be rescheduled. The simulation-based methods can be time-consuming and require a series of appropriate assumptions on the interactions of trains, infrastructures, and virtual interruptions. There is a lack of algorithmic solutions to identify secondary disruptions and effective recovery interventions based on empirical evidence.

Table 1 shows a comparison of recent disruption detection studies in metro systems and illustrates the contribution of our research. We conclude this section with a summary of gaps identified in the literature:

1. For the detection of service (operational) disruption in metro systems, methods based on incident logs and social media data can be unreliable due to inevitable human errors and missing observations. Recent SCD-based methods capture abnormal passenger behaviour as an indicator of disruption occurrence, but they might not detect service disruptions due to the lack of one-to-one association between service delays and abnormality in passenger behaviour.
2. In metro systems, current detection methods rarely pay attention to the propagation of disruption across space (along the network) and time. There is no clear understanding of primary and secondary disruptions. Neither the SCD nor the social media data contain effective information to identify the disruption propagation.

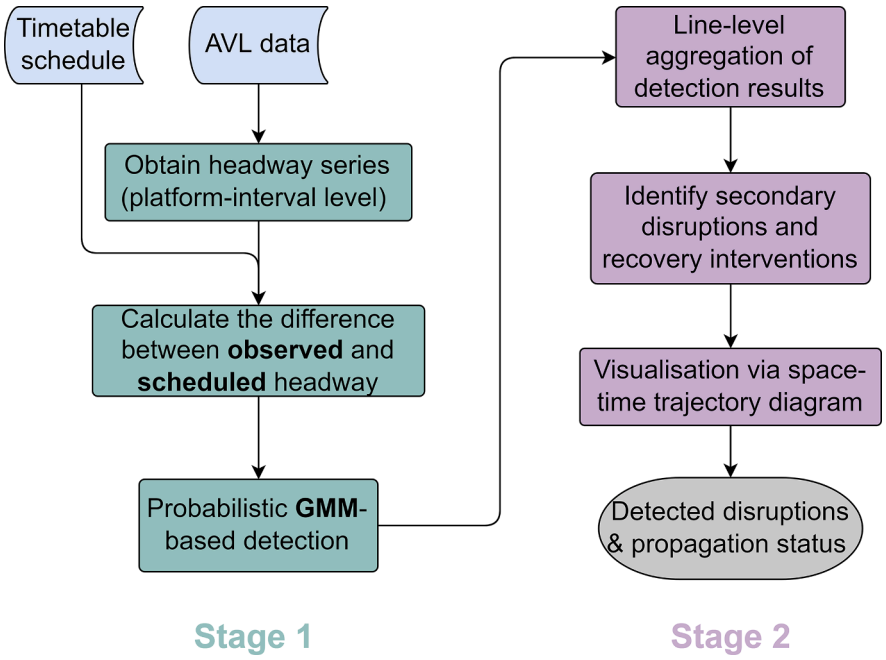

Fig. 2. Flowchart of the methodological framework.

3. For metro systems, there is a lack of empirical evidence for the quick identification of effective interventions.

In this paper, we show that both research gaps can be addressed by automatically detecting service disruptions using large-scale automated vehicle location (AVL) data.[6] We use the deviation in headways from the scheduled ones as an indicator of disruption occurrence. Moreover, by merging the disruption detection results with train movement trajectories, spatiotemporal propagation of service disruptions is identified. Analysis of the disruption status of stations on a metro line over time provides information about the origin of the service disruption and the extent of spill over effects. Such information will be vital in devising operational countermeasures.

3. Methodology

Our detection approach has two stages. First, in Section 3.1, we demonstrate how Gaussian mixture models (GMMs) can be applied to probabilistically detect platform-level metro service disruptions. Second, in Section 3.2, we analyse the line-level disruption propagation to identify the primary source and secondary spread of the disruption. Fig. 2 details all steps of the proposed detection framework.

3.1. Stage 1: Probabilistic detection with Gaussian mixture models

This section describes GMMs and motivates their application in detecting abnormal headways. The train service analysis to extract the observed headways from the AVL data and scheduled headways from the timetable is presented in the first subsection. GMM specification, its parameterisation, and the maximum-likelihood estimation are described in the next subsection. The procedure of applying the GMM-based model to detect service disruptions is presented in then the subsequent subsection. Finally, the last subsection details a simulation-based algorithm to derive optimal thresholds of disruption probabilities to designate a station to be disrupted.

3.1.1. Train service analysis

In urban metro systems, train services are planned according to a timetable defined by operators. Headway, the inverse of train frequency or the distance between two successive trains measured in time or space, is the key measure of service quality.[7] Under regular operating conditions, the observed headway is similar to the scheduled headway with some natural deviation. However, when

[6] In general, AVL data is owned by metro operators and can be provided in real-time or stored as historical records for offline use. Researchers may request access to these data and developed approaches can be implemented by metro operators.

[7] In this research, we define time as the unit of measurement, so headway here represents the tip-to-tip time from the departure of one train to the departure of the next train on a platform.

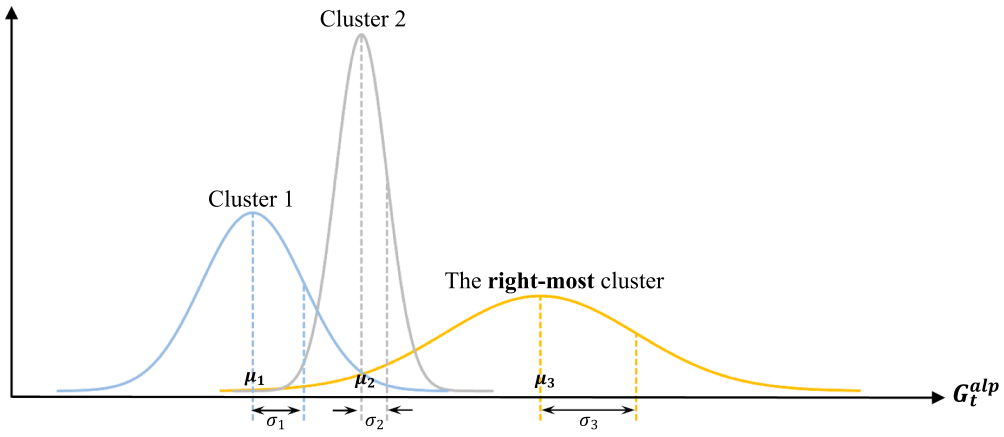

Fig. 3. An illustration of the right-most cluster.

train services are interrupted, the difference between the observed and scheduled headway is likely to exceed an acceptable level. Thus, abnormal (overlong) headway can be regarded as an indicator of the service disruption occurrence.

As shown in Fig. 2, the observed headway series (denoted by H) are extracted from the AVL data for each platform. The scheduled headway series (S) are obtained from service timetables. We define the deviation between observed and scheduled headways by

$$\boldsymbol{G}_{dt}^{alp} = \boldsymbol{H}_{dt}^{alp} - \boldsymbol{S}_{dt}^{alp}, \tag{1}$$

where vector $\boldsymbol{H}_{dt}^{alp}$ denotes the observed train headways on platform p ($p = 1,\cdots,P$) of line l ($l = 1,\cdots,L$) at station a ($a = 1,\cdots,A$) on a given day d ($d = 1,\cdots,D$) during time interval t($t = 1,\cdots,T$). The vector $\boldsymbol{H}_{dt}^{alp}$ stacks the headways of trains departing from platform p in time interval t. We use the same indices for $\boldsymbol{S}_{dt}^{alp}$ and $\boldsymbol{G}_{dt}^{alp}$. Vector $\boldsymbol{G}_{dt}^{alp}$ stacks the deviations between observed and scheduled headways in the time interval t on platform p. Please note that t segment the day into multiple predefined time intervals. The length of interval can be determined based on the magnitude and variation in scheduled headway on the given platform. One should specifically ensure that within each time interval, (i) there are adequate headway observations, and (ii) the corresponding scheduled headway remains close in the interval. On these grounds, we set the interval length of 30 min in the present case study (that is, $T = 36$ for 18 service hours). Then, we merge multiple days of observations. For the entire study period of D days, the platform-interval level headway deviation data is stacked in a vector as follows:

$$\boldsymbol{G}_{t}^{alp} = \left\{\boldsymbol{G}_{1t}^{alp}, \boldsymbol{G}_{2t}^{alp}, \cdots, \boldsymbol{G}_{Dt}^{alp}\right\}. \tag{2}$$

Thus, we estimate one detection model for each platform-interval to identify abnormal headway deviations with the input data $\boldsymbol{G}_{t}^{alp}$ across all days.

3.1.2. GMM and disruption identification

The GMM is a probabilistic model to identify subpopulations or clusters of observations with similar characteristics within a population (McLachlan and Basford, 1988; Peel and McLachlan, 2000). For example, in the context of this study, GMM can endogenously identify clusters of regular and abnormal headway deviations. There are two motivations behind using GMM to detect abnormal (overlong) headways. First, without true labels (normal and abnormal) on the headway deviation data, this detection problem is an unsupervised learning problem. Moreover, due to the nature of unexpected incidents or failures, the headway data is expected to contain relatively fewer anomalous observations (i.e., small subpopulation with abnormal characteristics). Since higher headway deviations indicate more severe disruption, such monotonicity can assist in naturally grouping even fewer abnormal headways into the right-most cluster (that is, with the highest cluster mean; see Fig. 3). Thus, GMM can address this unsupervised learning problem by systematically separating abnormal headways from other clusters of regular headways. Second, GMM is probabilistic, and thus we can obtain the probability of each headway observation to belong to the right-most cluster. In other words, the GMM-based detection method provides the probability of a platform being disrupted during a specific interval.

Compared to the deterministic detection methods based on empirical rules (e.g., three standard deviations away from the mean), the GMM-based method does not require the analyst to define subjective thresholds to characterise a headway to be abnormal. However, the threshold on the probability of a headway deviation belonging to the right-most cluster is required in the GMM-based model to identify the disrupted headways. Such thresholds can be learned through a semi-synthetic simulation (see **Selecting parameters through simulation** for details). The data-dependent probabilistic thresholds in GMM perform better than the subjective thresholds in deterministic models in detecting minor abnormalities because the former is normalised but the latter suffers from scale of standard deviation. For example, when the standard deviation of the observed headway is longer than 5 min, minor or moderate service interruptions (i.e., those under 10 min) cannot be identified using a deterministic rule in which headways beyond two standard

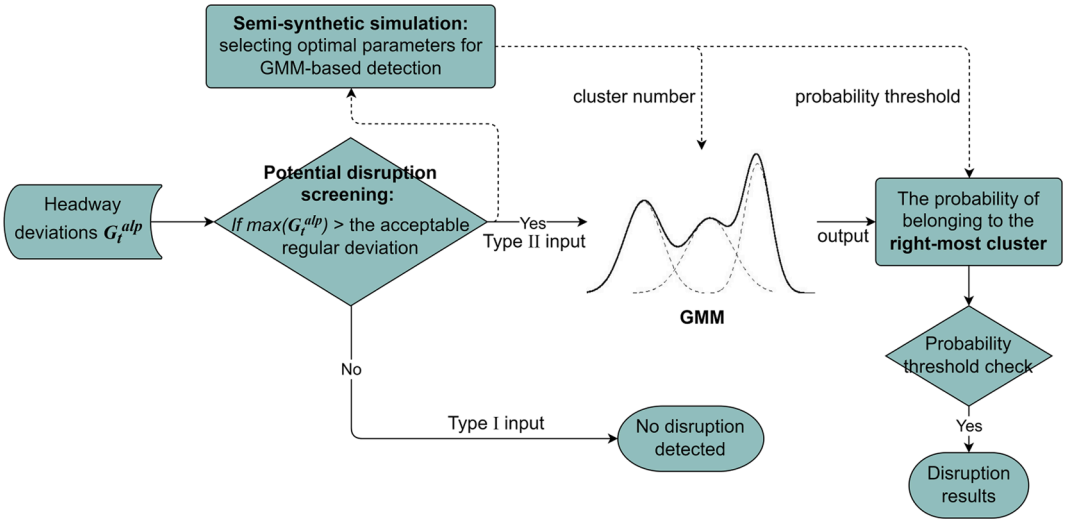

Fig. 4. The procedure of applying the GMM-based detections.

deviations of the mean are designated as outliers.

We succinctly discuss one-dimensional GMM formulation in the context of this study. A Gaussian mixture density of an observation G_{it}^{alp} is a weighted sum of M component densities:

$$p\left(G_{it}^{alp}\right) = \sum_{j=1}^{M} w_j\, p_j\left(G_{it}^{alp}\right), \tag{3}$$

where $G_{it}^{alp}(i = 1, \cdots, N)$ is an observation of vector $\boldsymbol{G_t^{alp}}$ (the headway deviations belonging to a specific platform-interval across all days), w_j is the mixture weight of the j^{th} component, and $p_j()$ is the Gaussian density of the j^{th} component with mean μ_j and variance σ_j^2:

$$p_j(G_{it}^{alp}) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{\left(G_{it}^{alp} - \mu_j\right)^2}{2\sigma_j^2} \right) \tag{4}$$

The mixture weights satisfy the following conditions:

$$\sum_{j=1}^{M} w_j = 1 \ \text{ and } \ 0 \leq w_j \leq 1 \tag{5}$$

The log likelihood function of observation for platform-interval can thus be written as:

$$\log p\left(\boldsymbol{G_t^{alp}}\right) = \sum_{i=1}^{N} \log\left\{ \sum_{j=1}^{M} w_j p_j\left(G_{it}^{alp}\right) \right\} \tag{6}$$

Here $\left\{\mu_j, \sigma_j, w_j\right\}_{j=1}^{M}$ are the identified parameters in GMM, which are obtained by maximising the loglikelihood presented in Equation (6). Since direct maximisation of the loglikelihood is cumbersome, we resort to an expectation–maximisation (EM) algorithm to maximise the loglikelihood (Dempster et al., 1977; Bansal et al., 2018). The probability of the headway difference $\left(G_{it}^{alp}\right)$ belonging to the j^{th} component is obtained using the estimated parameters and Bayes rule.

3.1.3. The procedure of applying the GMM-based detection model

Fig. 4 displays the procedure of GMM-based detections. Note that the distribution of the right-most cluster (with the highest mean headway deviation) depends on the variation in input headway deviations $\boldsymbol{G_t^{alp}}$. Since disruptions do not occur often, composition of the headway deviation data for different platform-intervals can inherently be of two types – (i) all regular observations with the headway deviations close to zero; (ii) both normal and abnormal headway deviations. While training GMM with the first type of data, the mean and standard deviation of the right-most cluster is likely to be small (e.g., zero to one minute). However, since we always focus on the right-most cluster to identify the service disruptions, the right-most cluster with a narrow tail and negligible mean headway deviation can be wrongly identified as a cluster of disrupted instances. GMM is likely to perform well for the second type of data, but we also want to circumvent the false detection of abnormal headways (that is, disruptions) in the first type of data.

To avoid the false detection of disruptions, we first check whether the input data have potential disrupted observations (type II) or

not (type I). If all headway deviation data is lower than the maximum acceptable regular deviation, we conclude that the platform experiences no disruption during a specific interval and therefore GMM estimation is not required. This concludes the detection process for type I input. Conversely, if there are headway deviations above the acceptable threshold, such input is categorised as type II with potential disruptions and the GMM detection step is applied on it. It is worth noting that the maximum acceptable regular deviation for each platform during a specific interval depends on the scheduled headway. For instance, if the scheduled headway at a metro platform in peak hours is 2 min, then service delays of approximately 1 to 2 min are acceptable, but 10-minute delays are not. On the other hand, at another platform, where the scheduled headway is 20 min, the delay of 5 to 10 min may be treated as an acceptable headway deviation. The acceptable headway deviation also depends on the metro operator's aspirations to provide a reliable service. We discuss the selection of acceptable headway deviations in Section 5.1.

In addition to the probability threshold for the right-most cluster, the number of clusters (M) needs to be selected to apply GMM for disruption detection. The number of subpopulations is often selected based on the Silhouette score and Bayesian information criterion (Lord et al., 2017), but in the context of this research, the choice of the number of clusters should not depend on statistical criteria. Specifically, we should ensure that the right-most cluster only contains the potential anomalies to avoid or minimise false disruption detections (see Fig. 3). Thus, the number of clusters and the optimal threshold on the probability of a headway deviation belonging to the right-most (abnormal) cluster are selected using semi-synthetic simulations. The simulation design is presented in the next subsection.

3.1.4. Selecting parameters through simulation

Instead of subjectively selecting the probability threshold and the number of clusters, we adopt a simulation-based grid-search method. The main idea is to use the empirical distribution of the headway deviation data to simulate new data and label a certain proportion (e.g., 1–5%) of the simulated headway deviation as disrupted. This proportion depends on the ratio of the observed headway deviations that exceed the maximum acceptable threshold for regular deviations. The data-generating process can be changed by varying the percentile of empirical deviation data used for the simulation and proportion of disrupted headway deviation (see Section 5.2 and Appendix A for details). Based on the simulated dataset with labelled disruptions, the detection problem can be translated into a supervised learning problem and the GMM's prediction accuracy can be tested under different combinations of the number of clusters and probability threshold. The proposed model selection procedure consists of the following steps, which will be run once for each platform-interval:

i). Derive a sample from the empirical cumulative distribution function of the observed headway deviations for a platform-interval, to generate the undisrupted simulation dataset. Calculate the proportion of potential abnormal deviations (that over the maximum acceptable threshold for regular deviations) in the given type II input data, and use this proportion to generate labelled disruptions.

ii). Run GMM-based detection models on the simulated headway deviation data for different number of clusters (e.g., ranging from 2 to 20) and threshold probabilities (e.g., $\{0.99, 0.98, 0.97, \cdots, 0.75\}$).

iii). For each combination of the number of clusters and probability threshold, compute performance measures: precision, recall, F1 score and accuracy.[8] Precision is the ratio of correctly detected disruptions to the total detected disruptions. Recall is the ratio of correctly detected disruptions to all the labelled disruptions. F1 score is the weighted average of precision and recall. Accuracy is the ratio of correctly detected observations to the total observations. To mitigate the simulation noise, repeat step (i) and (iii) 1000 times and obtain the average value of performance measures.

iv) Now create a two-way table of performance measures with rows indicating the number of clusters, and columns indicating the optimal threshold probability and the corresponding average value of performance measures. This two-way table is used to identify the optimal number of clusters.

v) Finally, select the best combination of the two parameters (cluster number and threshold probability) obtained in step (iv), and use them to conduct GMM detection on the actual observed deviation data.

3.2. Stage 2: Secondary disruption and recovery intervention identification

The GMM-based detection method provides information on the location, time and duration of disruptions. In this section, we use this output to find the linkages between the detected disruptions at consecutive platforms along a metro line. We categorise disruptions into two types: primary disruption and secondary disruption. Primary disruption means that the service interruption is *originated* at the given platform during a specific period. In contrast, secondary disruption at a platform is caused by a primary disruption at one of the upstream or downstream[9] platforms along the metro line. As discussed earlier, this is the first study for metro systems that provides an effective algorithmic solution to identify primary/secondary disruptions and recovery interventions based on empirical data. Since metro systems reboot every day, we analyse metro line operations on a specific day of disruption(s) using the following steps:

[8] Formulas of the performance measures: $Precision = \frac{True\ \ positives}{True\ \ positives+False\ \ positives}$, $Recall = \frac{True\ \ positives}{True\ \ positives+False\ \ negatives}$, $F1\ \ score = \frac{2\times Recall\times Precision}{Recall+Precision}$, $Accuracy = \frac{True\ \ positives+True\ \ negatives}{True\ \ positives+False\ \ positives+True\ \ negatives+False\ \ negatives}$.

[9] Train delays may spread in the opposite direction, upstream to the disrupted station as well, due to queueing or dispatching interventions.

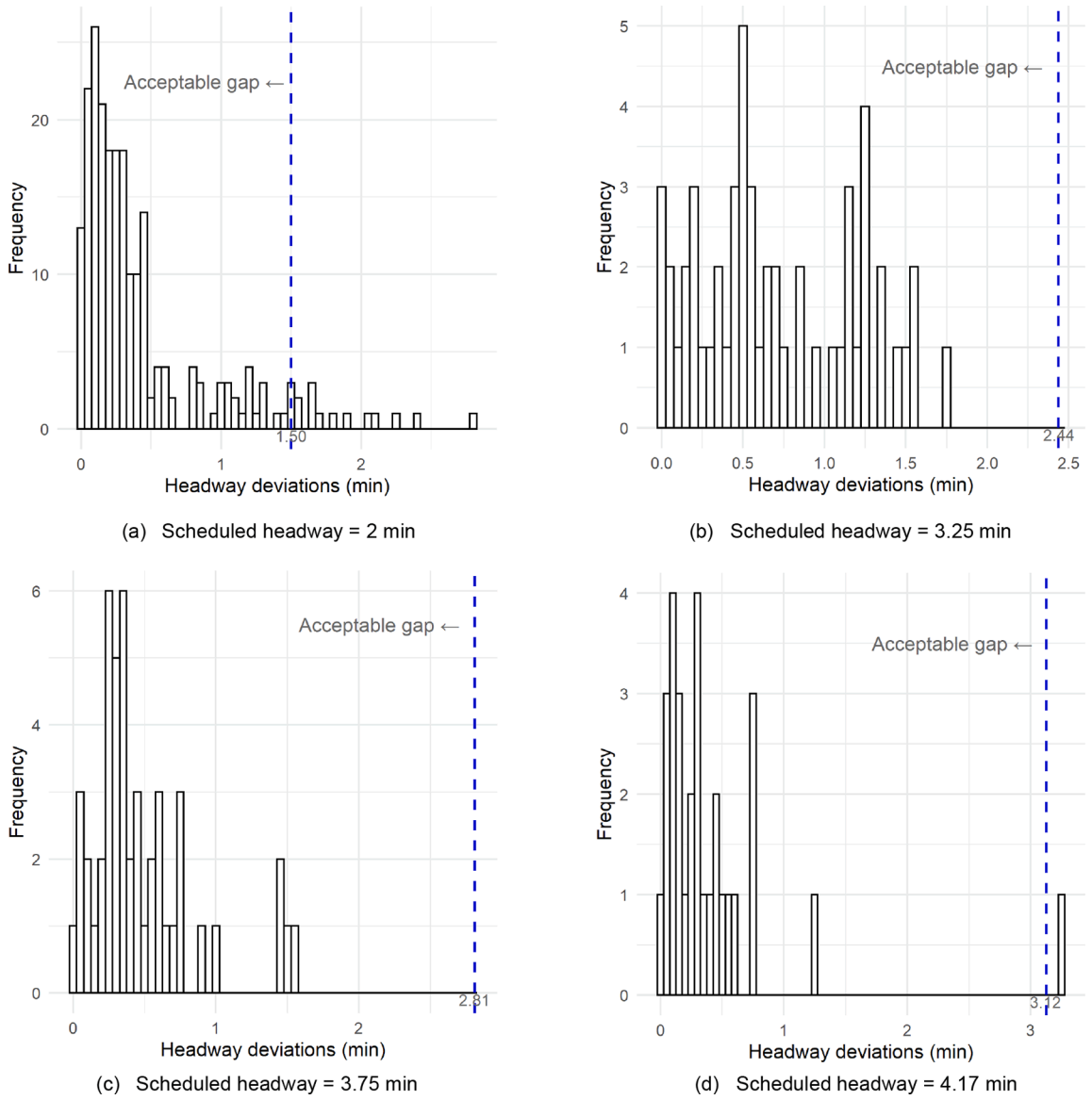

Fig. 5. The histogram of observed headway deviations for different scheduled headways from the given platform-interval of the example station.

i) Pool the GMM-based detection results on all platforms of a metro line (with the same direction of train services) on the specific day.

ii) Sort all disruption records based on the start time of disruptions. Mark the first record as a *primary disruption*. For the next record, if (1) the platform is downstream to the primary disruption location (follow the train direction); (2) the start time of the disruption is slightly later; and (3) the train ID and trip ID are the same, this record is marked as a *secondary disruption*. However, for the next record, if the platform is downstream to a primary disruption location (follow the train direction) and the start time of disruption is slightly later, but the train ID and trip ID are not the same, this record is marked as a *secondary disruption with an intentional dispatching intervention* from the operator. Such interventions aim to restore normal services and reduce the impact of delays on passenger waiting time. Repeat this process until the upcoming record breaks the spatiotemporal continuity of start time and downstream location conditions.

iii) Repeat step (ii) until all disruption records are marked as either primary disruption, secondary disruption, or secondary disruption with intentional dispatching intervention.

iv) Merge the daily disruption records obtained in step (iii) with the corresponding train trajectory data to visualise disruptions and the train movement using the space–time diagram.

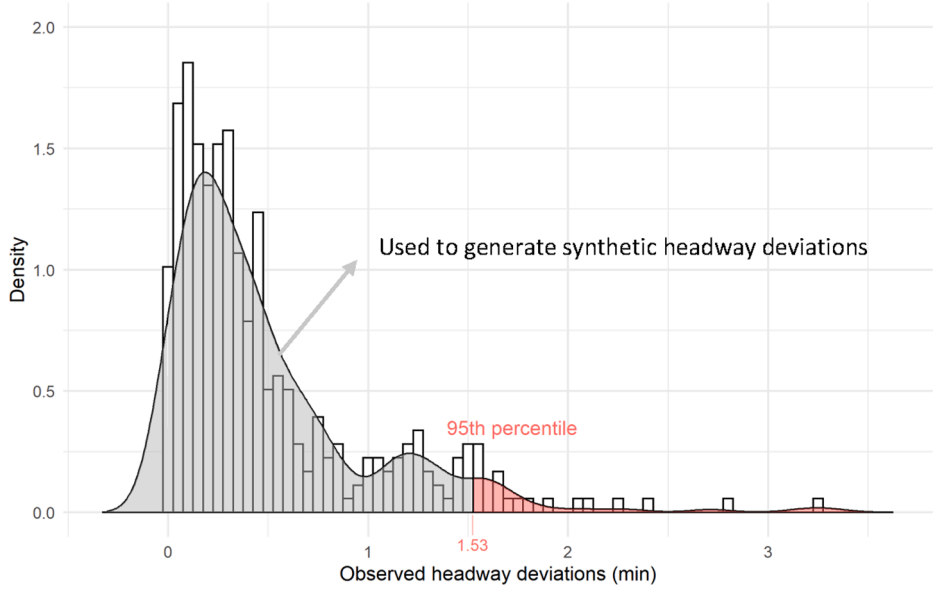

Fig. 6. Histogram of actual headway deviations observed from the given platform-interval.

4. The case study

In this paper, to illustrate the detailed process of disruption detection, we select a densely used line in a major metro system in Asia to carry out the case study. The selected metro line has 16 stations with tracks of a total length of 16.9 km. Being the link between central business areas and suburban areas, this line is constantly busy and crowded. We detect service disruptions that occurred in both upward and downward directions. During the study period (54 weekdays from 01/01/2019 to 31/03/2019, excluding holidays and days of incomplete data), the scheduled headway of the given metro line ranges from 2 to 10 min with an average of around 3 min. We choose 30 min as the interval to group the headway data of each platform, which also ensures that each platform-interval group has sufficient headway observations.[10] The daily service time of the metro system starts at 6:00 and ends at 24:00, which is thus divided into 36 intervals to conduct GMM-based detections. Therefore, taking account of 16 stations with two platforms and 36 intervals for each platform, our dataset is aggregated into a total of 1152 platform-interval groups. We also identify secondary disruptions and the effective recovery intervention from metro operators. Finally, the information on the detected disruptions is collected to build a high-quality database of service interruptions.

The following data are used to detect and evaluate the service disruptions. We conducted data processing and analysis using open-source R software (version 4.1.1).

Automated vehicle location (AVL) data: The AVL data from 01/01/2019 to 31/03/2019 are provided by the metro operator. Public holidays including the New Year and the Spring Festival are excluded. We consider this duration as our study period. The AVL data contain information on train ID, trip ID, the timestamp of train movements (including precise departure and arrival times), and the location of train movements (including station, line and directions). The resolution of time stamps exacts to one second. By using the AVL data, we can extract headway series from the consecutive train movements on each platform.

Timetable schedules: The scheduled arrival and departure time of train services on the selected line, provided by the metro operator. We utilise this information to extract scheduled headways.

Incident logs: The manual inspection record of incidents, including information such as occurrence time, location, cause and duration of disruptions, provided by the metro operator. Incident logs are used to validate our detection results.

Pseudonymised smart card data (SCD): The SCD contain information on the time and location of tap-in and tap-out transactions throughout the system, recording individual trips. In this research, the role of SCD is constrained to illustrating the limitations of the demand-based disruption detection methods.

5. Results and discussion

The results are presented in four steps. First, we illustrate the method of screening input data and how to identify platform-intervals where the possibility of disruptions cannot be excluded. Second, we showcase the simulation approach to the optimisation of key hyper-parameters of the proposed GMM method. The third subsection presents the final outputs of the GMM disruption detections. The

[10] The observed headway is more consistent within each 30-minute interval. However, due to the various frequency of service at different times of the day, the mean headway is longer during off-peak 30-minute intervals and shorter during peak 30-minute intervals.

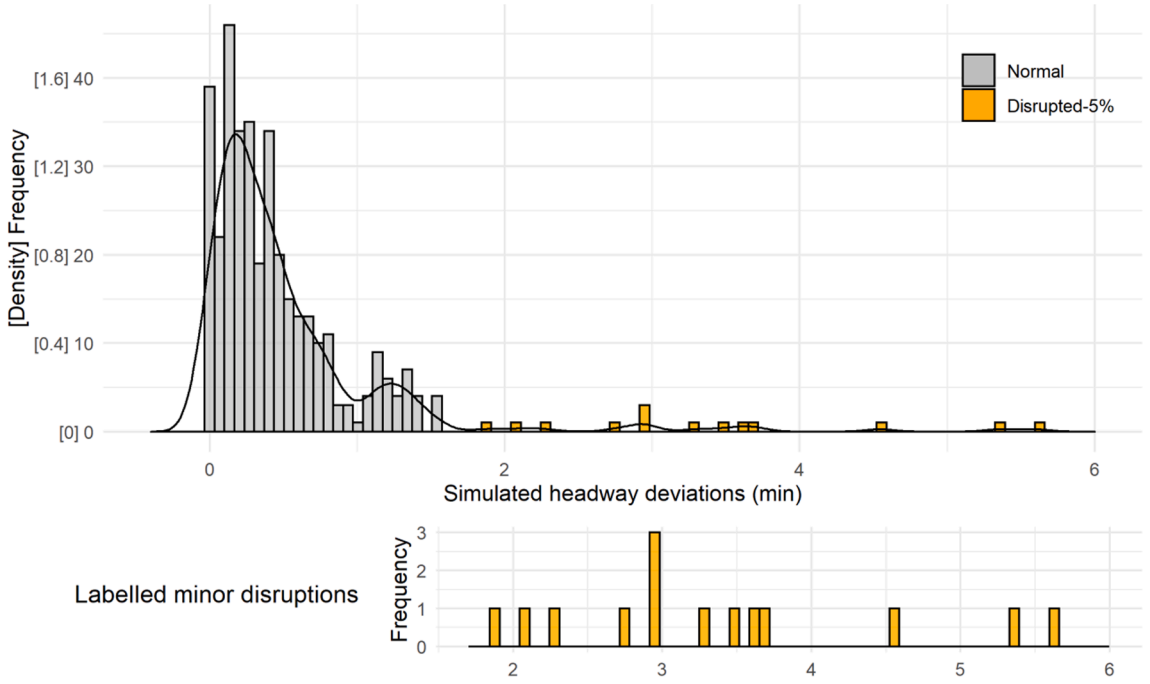

Fig. 7. Histogram of a sample of the simulated headway deviations for the given platform-interval. The "labelled minor disruptions" means the simulated headway deviations that are marked as disruptions and the duration of such disruption is relatively short (2 to 6 min).

results in the first three steps are presented through a sample dataset from a randomly selected station of the studied metro line, in the initial stage of the morning peak period. Finally, Section 5.4 demonstrates disruption propagation through the identification of secondary disruptions and dispatching interventions along the entire line.

5.1. *Input data check: Screening potential disruptions (Type II)*

At the example station, northbound, 7:30–8:00 a.m., we first check the input observed headway deviation data. The scheduled headways on this platform-interval range between 2 and 4 min. To guarantee the reliable service of early peak hours in the morning, we determine that the acceptable headway deviations should be lower than 75 % of the scheduled headway. This threshold is set arbitrarily, based on the intuition that if no train is delayed by more than another scheduled headway, including a 25 % safety gap, then it is very unlikely that significant disruptions happened within the 30-minute interval. Fig. 5(a) to 5(d) display the histogram of the actual headway deviations under different scheduled headways. The dashed lines represent the 75 % boundaries defined above. In plots 5(a) and 5(d), there are observed deviations above the acceptable level, which means that the input data is type II and we cannot exclude the presence of disruptions. Therefore, we proceed to the next step of our analysis.

5.2. *Optimal number of GMM clusters and probability threshold*

Before applying detection models, we run semi-synthetic simulations to obtain the optimal values of two critical parameters: the number of clusters in our GMM approach and the probability threshold above which observations in the right-most cluster indicate a disruption. The actual headway deviations are used to generate the synthetic simulated datasets. Fig. 6 shows the overall distribution of observed headway deviations regardless of their schedules. The graph shows that 95 % of the actual deviations are less than 1.5 min.

Following the procedure in Section 3.1.4, we first generate a synthetic dataset of *undisrupted* headway deviations, which is drawn from the empirical distribution of the observed deviations truncated at the 95th percentile. We ensure that the sample size of the synthetic data matches with that of the empirical data. Subsequently, a certain proportion of disruptions are generated based on a lognormal distribution. The mean of the simulated lognormal distribution μ_{syn} is set according to the scheduled headway,[11] and the standard deviation σ_{syn} is set to achieve the varied lengths of disruptions. These disruptions are then randomly allocated and added to the original synthetic deviations, thus forming the *disrupted* headway deviations.

[11] The mean of (lognormally distributed) simulated disruption durations is set as a multiple of the (log) scheduled headway, where the multiplier is derived from the available incident log data. We compute the mean and standard deviation of the actual distribution from the incident log records, compute the same for the simulated lognormal distribution as a function of the multiplier, and solve for the multiplier by minimising the difference between the two moments of the distributions. Within each platform-interval, the multiplier is selected based on this automated procedure.

Table 2
The two-way table of simulation performance and optimal GMM parameters.

Number of clusters	Precision	Recall	Accuracy	Optimal threshold
2	0.152	0.998	0.725	1.000
3	0.176	0.996	0.769	1.000
4	1.000	0.813	0.991	0.933
5	1.000	0.871	0.994	0.958
6	1.000	0.896	0.995	0.960
7	1.000	0.907	0.995	0.965
8	1.000	0.923	0.996	0.974
9	1.000	0.930	0.997	0.982
10	1.000	0.931	0.997	0.980
11	1.000	0.914	0.996	0.975
12	1.000	0.937	0.997	0.984
13	1.000	0.946	0.997	0.994
14	1.000	0.943	0.997	0.991
15	**1.000**	**0.947**	**0.997**	**0.994**
16	1.000	0.912	0.996	0.979
17	1.000	0.903	0.995	0.968
18	1.000	0.890	0.995	0.959

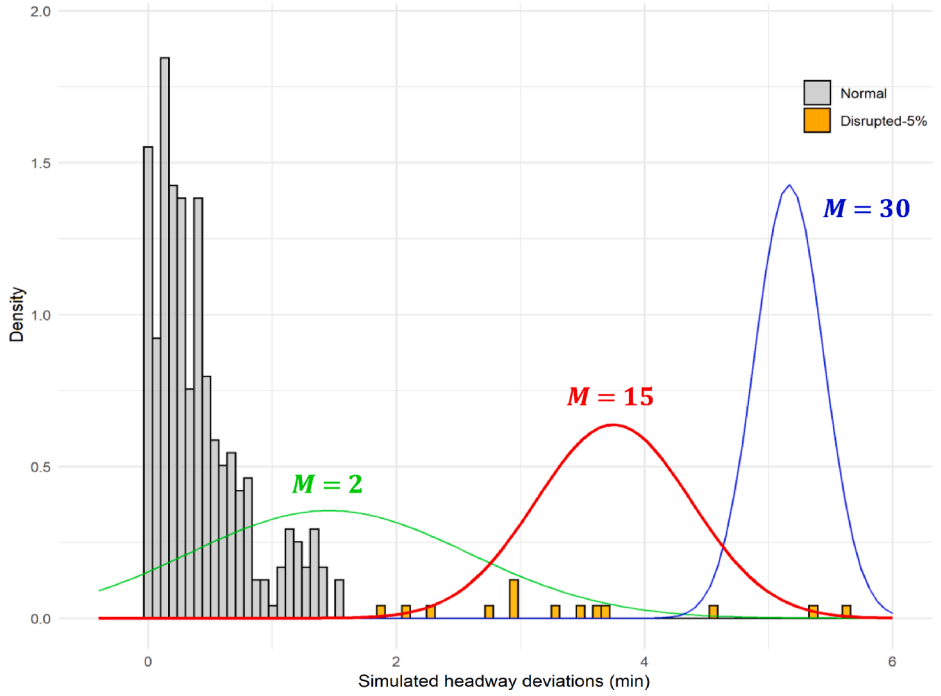

Fig. 8. Changes in the right-most clusters of the estimated GMM, under different number of clusters (M). Each solid line represents the distribution of the right-most cluster for a given cluster number.

In the present example, 5 % of the observed headway deviations are above the acceptable level. We choose this proportion to generate disruptions. For the distribution of disruption durations, the μ_{syn} is set as 1.2 times the natural logarithm of the scheduled headway, and the standard deviation σ_{syn} is set to be 0.3. The disrupted synthetic deviations range between 1.5 and 8 min. Fig. 7 displays the empirical distribution for one sample of the synthetic simulated dataset. The grey bars represent undisrupted headway deviations while the orange bars indicate labelled disruptions. In the Appendix, we perform robustness checks regarding the chosen percentile for the sampling of undisrupted observations. Also, we perform sensitivity analyses for different platform-interval combinations and different rates of simulated disruptions.

With pre-defined labels of service status, semi-synthetic simulations enable us to transform the disruption detection task into a supervised learning problem. Based on the simulated headway deviations, we obtain the optimal GMM parameters by comparing the detection performance under a wide range of possible combinations of the cluster number and the probability threshold.

Table 2 summarises the three performance measures (precision, recall, accuracy) and the optimal probability thresholds for the given number of clusters. Due to limited space, the cluster numbers shown range from 2 to 18 (each row represents the average result of

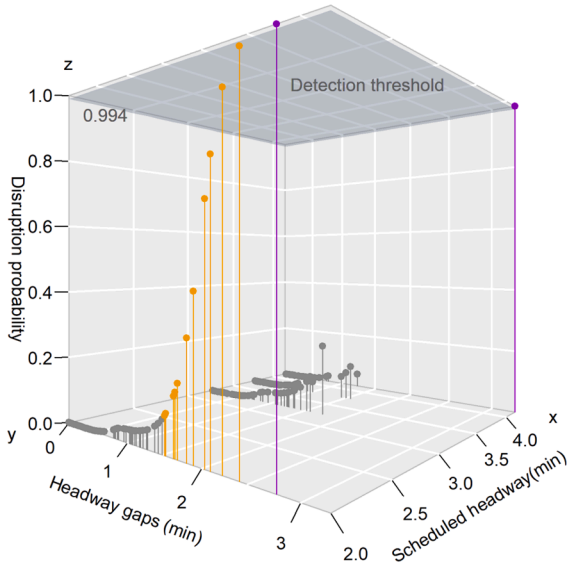

Fig. 9. The final detection results with the probability threshold of the example station (northbound, 7:30–8:00 a.m.).

Table 3
Performance benchmarking results (simulation - averages of 1000 runs).

Detection methods	Threshold	Precision	Recall	F1-score	Accuracy
Simple comparison with given thresholds	2 min	0.664	0.998	0.798	0.976
	5 min	1.000	0.140	0.257	0.959
Deterministic detection based on the empirical rule	Mean + SD*	0.564	0.999	0.713	0.961
	Mean + 2SD*	1.000	0.821	0.899	0.991
	Mean + 3SD*	1.000	0.641	0.778	0.983
GMM-based detection	Optimal probability threshold	1.000	0.947	0.972	0.997

* SD denotes the standard deviation of the observed headway deviations.

1000 simulation runs). We find that, when the number of clusters is set to 15 and the probability threshold is 0.994, both the detection precision rate and overall accuracy reach their maximum values (above 0.997). The balance between the precision and recall rate also reaches the best. Fig. 8 shows how the right-most cluster changes under different choice of cluster numbers. As the GMM clusters increase from 2 to 30, the right-most cluster gradually shifts to the right of x-axis with higher mean and lower standard deviations. Meanwhile, the probability of all disrupted headway deviations belonging to the right-most cluster keeps increasing until the number of clusters reaches 15. When the cluster number continues to grow, such probability starts to drop as the less spread right-most cluster tends to cover fewer disruptions. Thus, in the formal GMM-based detections, the optimal 15 clusters and the 0.994 probability threshold are applied for the given platform-interval in this example.

5.3. GMM detection results

The GMM with optimal parameters is then applied to the actual headway deviation data to derive final detection results of the above example. Fig. 9 presents the detection results in the form of a three-dimensional plot. The y-axis of the 3D plot represents observed headway deviations in the given platform-interval, the x-axis represents the corresponding scheduled headways, and the z-axis represents the probability of belonging to the right-most cluster (refer to Section 3.1.3). The colour of scatter points indicates detection decisions. The grey points refer to normal headway deviations that are within the corresponding acceptable levels. Their disruption probabilities are less than 22 %. The yellow points tend to include all possible outliers, with the disruption probability ranging from 10 % to 99 %. To achieve the highest detection accuracy, we rely on the optimised probability threshold. In this case, only two observations (highlighted in purple) are above 0.994, and they are finally identified as disruptions.

In terms of the entire line, we compare our detection results with manual incident logs from the metro operator. For medium to severe interruptions that are between 5 min and several hours long, all reported disruptions have been detected by the proposed GMM method. For minor service interruptions that range from 2 min to 5 min, 96 % of them have been successfully identified. The data-driven detection also provides more supplementary results that may have been omitted in human inspections. To detect disruptions in a 30-minute platform-interval, the running time on a laptop with 2.80 GHz CPU and 16 GB RAM is less than 5 min. When the proposed method is implemented on a high-performance computing facility with parallel implementation, it would be suitable for monitoring large metro networks under the expected information refresh rate.

Table 4
A sample of selected disruption records with the corresponding identification results.

Disruption ID	Start time	Station ID	Train ID	Duration (min)	Category
18	10:48:44	2	42	00:09:48	Primary
19	10:52:03	5	53	00:04:33	Intervention
20	10:52:05	3	42	00:10:09	Secondary
21	10:55:57	4	42	00:08:22	Secondary
22	10:59:41	5	42	00:06:58	Secondary
23	11:02:06	6	42	00:06:35	Secondary
24	11:03:55	7	42	00:06:34	Secondary
25	11:05:05	11	53	00:04:59	Intervention
26	11:05:50	8	42	00:06:38	Secondary
27	11:07:30	9	42	00:06:46	Secondary
28	11:07:53	12	53	00:05:09	Intervention
29	11:09:14	10	42	00:07:02	Secondary
30	11:09:51	13	53	00:05:09	Intervention
31	11:11:32	14	53	00:05:08	Intervention
32	11:13:09	11	42	00:04:57	Secondary
33	11:13:33	15	53	00:04:58	Intervention
34	11:14:44	16	53	00:04:58	Intervention
35	11:16:07	12	42	00:04:46	Secondary
36	11:18:05	13	42	00:04:47	Secondary
37	11:19:45	14	42	00:04:42	Secondary
38	11:21:36	15	42	00:04:47	Secondary
39	11:22:46	16	42	00:04:47	Secondary
40	13:32:14	2	40	00:04:04	Primary
…	…	…	…	…	…
118	20:24:10	2	70	00:06:49	Primary
119	20:26:17	4	44	00:04:43	Intervention
121	20:27:38	3	70	00:07:12	Secondary
122	20:28:10	5	44	00:04:58	Secondary

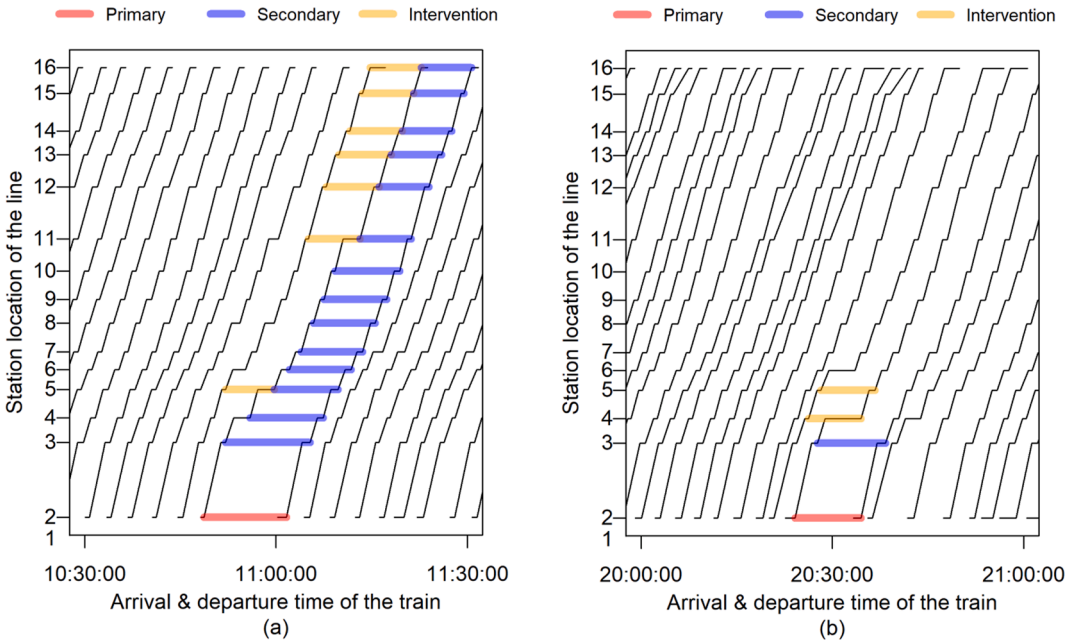

Fig. 10. Spatial-temporal train movement diagram with detected disruptions.

As for the validation via simulated detections, in all stations of the given line under both minor and mixed disruption scenarios,[12] the average detection accuracy is above 0.99. Specifically, the average precision is nearly uncompromised, and the average recall rate is over 0.9. The corresponding robustness checks and sensitivity analyses, both based on semi-synthetic simulations, are demonstrated in the Appendix.

Using the simulated datasets of the above example with minor interruptions, we benchmark the performance of the GMM-based detection against two simpler alternative methods: predetermined delay thresholds and deterministic detection based on empirical rules. Results in Table 3 show that the proposed approach outperforms these alternative detection methods. As for the deterministic detections, the choice of subjective thresholds affects the detection accuracy significantly. An overly short threshold can increase the false positive rate by up to 44 % and an excessive threshold of 5 min can lead to 86 % of disruptions not being identified. Compared with the best performing deterministic threshold (mean + 2SD), the use of probabilistic threshold increases the recall rate by 15 % and raises the overall accuracy to 0.997.

5.4. Secondary disruptions and recovery interventions

In this subsection, we demonstrate how to apply the algorithm presented in Section 3.2 to identify secondary disruptions. Since the metro system closes after midnight and reopens the next morning, the identification is implemented on a daily basis. After the first step of pooling disruptions into the level of the entire line and partitioning based on date, Table 4 shows a sample of the detected disruptions during two off-peak periods (10:30–11:30 and 20:00–21:00). Considering the average scheduled headway of the line is around 3.5 min in these periods, we focus on detected disruptions over 4 min.

In the second step we sort these detection records by start time, as shown in Table 4. Fig. 10 visualises the disruptions and their categories with the corresponding train trajectory data in a space–time diagram. The first record (Disruption 18) is marked as the initial primary disruption. When moving to the next record, Disruption 19 starts slightly later than the primary-one (within the regular time of a full journey) and the platform location is downstream, but their train IDs are not the same. Thus, this record is marked as a secondary disruption with an intentional dispatching intervention from the operator. Then, moving to the third record, compared with the primary-one, Disruption 20 satisfies all three conditions of a secondary disruption; it starts later, at a downstream station, with the same train ID. We repeat the above procedure until we encounter a new record that breaks the temporal and spatial proximity. For instance, after Disruption 39, the location of Disruption 40 is once again at Station 2 and it occurs nearly-two hours later. In this case, Disruption 40 will be marked as a primary disruption again. We repeat the screening steps until all the records are processed.

The identification results within the two sample periods are shown separately in Fig. 10. The horizontal axis represents the arrival and departure time of trains at each platform, while the vertical axis shows the location of and distance between the stations along the line. The black solid lines are the trajectories of train movement, and the bold lines are detected disruptions. We confirm that both detections and secondary identification results match well with train trajectories, which is in line with the law of interruption propagation. For example, in Fig. 10(a), a primary disruption occurred at 10:48:44 at Station 2. On the one hand, this disruption spreads downstream along the line until the terminal station. On the other hand, metro operators act promptly at Stations 4, 5, and 11 to increase the dwell time of the last train preceding the disrupted one, thus avoiding further bunching effects. Due to these interventions, after Station 5, the disrupted train does not accumulate further delays. Similarly, in Fig. 10(b), a primary disruption occurred at 20:24:10 at Station 2. Then, immediate interventions take place at Stations 4 and 5 to slow down the previous train and relieve the delayed one from excessive passenger load. No further delays are identified after the primary disruption spreading to Station 5, and the train services return to normal.

This visual analysis indicates that the proposed detection framework is valid and highly effective for identifying disruptions and their categories. Furthermore, Fig. 10 also illustrates how identifying secondary disruptions can contribute to practical metro operations. In the space–time diagram of train movements, by labelling the secondary disruptions due to interventions, operators can easily locate the effective interventions used for mitigating delays, such as adjusting the dwell time of upstream trains. More importantly, with automated disruption classification, the operator can disentangle the frequency and severity of primary disruptions from subsequent time loss due to delay propagation and dispatching measures. This information is essential for both preparing recovery plans and analysing the resilience of metro systems.

6. Conclusions and future work

Service disruptions cause various challenges in urban metro systems, including delays, crowding, and declining passenger satisfaction. Operators need to monitor disruption occurrences closely in order to reduce their detrimental effects. With accurate information on the location, time, duration, and propagation process of disruptions, they can comprehensively assess the reliability and resilience of metro systems. Thus, the detection of service disruptions is a prerequisite of any further research on disruption management.

This research proposes a novel, probabilistic, unsupervised clustering framework to quantify the probability of an observed train headway being identified as abnormal. In contrast to traditional manual inspections and other detection methods based on social media data or smart card data, which suffer from human errors, limited monitoring coverage, and potential bias, our approach uses

[12] The simulated minor disruptions range from 1.5 to 8 min. The mixed disruptions are referred to as a mixture of minor interruptions and severe interruptions (over a few hours).

information on train trajectories derived from automated vehicle location (train movement) data. The proposed GMM approach assumes that the observed headway distributions are composed of a disrupted and multiple undisrupted subcomponents, where disruptions belong to the right-most subcomponent with the highest mean headway deviation. Our approach estimates the probability that a headway deviation observation belongs to the right-most cluster. We develop a simulation algorithm to infer the threshold probability, above which the headway observations are classified as disruptions. Finally, we extend the detection framework from the platform level to entire metro lines. We distinguish three categories of service delays: primary disruptions, secondary delays of the disrupted train at downstream stations, and delays of other trains due to dispatching interventions. To the best of our knowledge, this is the first study in the metro field which identifies secondary disruptions and the operator's effective recovery interventions using empirical data and algorithms.

The proposed method is applied in a case study of a densely used line in the given metro system. This illustrative application indicates that the detection accuracy of our method is very high. In all simulated scenarios for the entire line, the average precision is nearly uncompromised and the average detection accuracy is above 0.991. For minor service delays in the range of 1.5 to 8 min, the average recall rate is over 0.90. Even though the proposed simulation framework is based on simple assumptions and idealised conditions, these results highlight promising prospects for practical adaption.

Let us conclude this research by acknowledging some of the present limitations of the proposed method. In daily metro operations, service disruptions can be caused by unexpected infrastructure malfunctions (e.g., signal failures and track blockages), rolling stock breakdowns and accidents, planned maintenance works, or temporal dispatching adjustments. The fact that we cannot obtain the cause of service disruptions from automated train movement data is clearly a limitation of our data-driven detection method, as compared to manual data collection. Indeed, the main reason for this limitation is that we perform the detection only based on one data source. In line with this limitation, our future research will focus on merging more data sources to infer the cause of disruptions, such as manual incident logs, smart card data, news, and data from social media.

CRediT authorship contribution statement

Nan Zhang: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Daniel J. Graham:** Supervision. **Prateek Bansal:** Methodology, Writing – review & editing. **Daniel Hörcher:** Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors are grateful for the support of the Hong Kong MTR, the data provider of this research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the MTR. Prateek Bansal was supported by the Leverhulme Trust Early Career Fellowship, United Kingdom.

Appendix A. Robustness and sensitivity analysis

A.1. Robustness check against percentile of headway deviations in simulation

As mentioned in Section 5.2, the percentile of observed headway deviations, which we use to construct undisrupted observations in the simulation, affects the composition of the synthetic samples and detection accuracy. The higher the percentile is, the more overlap will be between the disrupted and regular headway deviations, potentially leading to lower detection accuracy. To evaluate the robustness of our GMM-based model in terms of this concern, we test for the following percentiles: 96, 97, 98, 99. The proportion of simulated disruptions in the sample data is kept at 5 %.

Table A1 summarises the changing trend of performance measures. Although the overall accuracy and recall rate continue to drop when the percentile increases, their minimum values are still over 0.99 and 0.93, respectively. Such results indicate that the proposed detection model is robust, even if the simulated deviations are generated from different percentile of empirical data.

Table A1
Robustness check against the choice of deviation percentile in simulation.

Percentile choice	5 % Disruptions			
	Precision	Recall	Accuracy	Optimal threshold
96 percentile	1.000	0.945	0.997	0.994
97 percentile	1.000	0.942	0.997	0.989
98 percentile	1.000	0.937	0.996	0.968
99 percentile	1.000	0.935	0.995	0.956

A.2. Sensitivity analysis of detection accuracy

Besides the mentioned example station (1) and time interval (7:30–8:00) in Section 5, we select another suburban station (2) of the studied metro line and an additional evening-peak interval (17:00–17:30) to make the validation more comprehensive. In validation, two types of disruptions are considered. The first type is minor disruption that follows log-normal distribution and ranges between 1.5 and 8 min. The second type also follows log-normal distribution but includes a wider range of headway deviations, consisting of both minor interruptions and more severe interruptions of few hours. The empirical distribution of the actual headway deviation is truncated at the 95th percentile, when simulating undisrupted observations. Under different combinations of stations, disruption types and disruption rates, the performance measures of detection are presented in Table A2.

We find that the detection performance varies slightly in each platform-interval, due to the different composition of headway deviation data. Overall, the detection is effective for both minor and mixed disruptions, under any given disruption rate.

Table A2
Results of sensitivity analysis for two example stations (1000 runs).

Disruption type	Station (Northbound)	Time	Disruption rate	Performance measures			
				Ave. precision	Ave. recall	Ave. F-score	Ave. accuracy
Minor interruptions (<8 min)	Example 1 (CBD area)	7:30–8:00	1 %	1	0.9441	0.9678	0.9995
			2 %	1	0.9453	0.9752	0.9989
			5 %	1	0.9465	0.9720	0.9974
		17:00–17:30	1 %	1	0.9999	1	1
			2 %	1	0.9999	0.9999	1
			5 %	1	0.9998	0.9999	1
	Example 2 (Suburb)	7:30–8:00	1 %	1	0.9033	0.9393	0.9992
			2 %	1	0.9074	0.9425	0.9983
			5 %	1	0.9157	0.9543	0.9960
		17:00–17:30	1 %	1	0.9823	0.9899	0.9998
			2 %	1	0.9839	0.9910	0.9997
			5 %	1	0.9870	0.9934	0.9994
Mixed interruptions (minor - few hours)	Example 1 (CBD area)	7:30–8:00	1 %	1	0.9934	0.9963	0.9999
			2 %	1	0.9931	0.9962	0.9999
			5 %	1	0.9928	0.9960	0.9999
		17:00–17:30	1 %	1	1	1	1
			2 %	1	0.9999	1	1
			5 %	1	0.9999	0.9999	0.9999
	Example 2 (Suburb)	7:30–8:00	1 %	1	0.9917	0.9950	0.9999
			2 %	1	0.9916	0.9960	0.9999
			5 %	1	0.9916	0.9956	0.9996
		17:00–17:30	1 %	1	0.9973	0.9985	1
			2 %	1	0.9973	0.9984	0.9999
			5 %	1	0.9972	0.9984	0.9999

References

Ashish, J., Wang, J., Yang, X., Ye, C., et al., 2018. TrackNet - A Deep Learning Based Fault Detection for Railway Track Inspection. In*: 2018 International Conference on Intelligent Rail Transportation, ICIRT 2018.* [Online].

Bansal, P., Daziano, R.A., Guerra, E., 2018. Minorization-Maximization (MM) algorithms for semiparametric logit models: Bottlenecks, extensions, and comparisons. Transport. Res. Part B: Methodol. 115, 17–40.

Bolla, R., Davoli, F., 2000. Road traffic estimation from location tracking data in the mobile cellular network. In: *2000 IEEE Wireless Communications and Networking Conference*. [Online]. 2000 pp. 1107–1112.

Briand, Anne-Sarah, Côme, E., Khouadjia, Mostepha, Oukhellou, L., 2019. Detection of Atypical Events on a Public Transport Network Using Smart Card Data. In *European Transport Conference 2019 Association for European Transport (AET)*. 1–3.

Büker, T., Seybold, B., 2012. Stochastic modelling of delay propagation in large networks. J. Rail Transp. Plann. Manage. 2 (1–2), 34–50.

Cano, J., Kovaceva, J., Lindman, M., Brannstrom, M., 2009. In: Automatic incident detection and classification at intersections. Volvo Car Corporation, pp. 9–245.

Carey, M., Kwieciński, A., 1994. Stochastic approximation to the effects of headways on knock-on delays of trains. Transport. Res. Part B: Methodol. 28 (4), 251–267.

Chen, J., Roberts, C., Weston, P., 2008. Fault detection and diagnosis for railway track circuits using neuro-fuzzy systems. Control Eng. Pract. 16 (5), 585–596.

Collins, C., Hasan, S., Ukkusuri, S.V., 2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. J. Public Transport. 16 (2), 2.

Corman, F., D'Ariano, A., Marra, A.D., Pacciarelli, D., Samà, M., 2017. Integrating train scheduling and delay management in real-time railway traffic control. Transport. Res. Part E: Logist. Transport. Rev. 105, 213–239.

D'Andrea, E., Marcelloni, F., 2017. Detection of traffic congestion and incidents from GPS trace analysis. Expert Syst. Appl. 73, 43–56.

D'Ariano, A., Pranzo, M., 2009. An advanced real-time train dispatching system for minimizing the propagation of delays in a dispatching area under severe disturbances. Networks Spatial Econ. 9 (1), 63–84.

de Bruin, T., Verbert, K., Babuska, R., 2017. Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks. IEEE Trans. Neural Networks Learn. Syst. 28 (3), 523–533.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) 39 (1), 1–22.

Dingler, M., Koenig, A., Sogin, S., Barkan, C.P., 2010. August). Determining the causes of train delay. In AREMA Annual Conference Proceedings.

Dollevoet, T., Corman, F., D'Ariano, A., Huisman, D., 2014. An iterative optimization framework for delay management and train scheduling. Flexible Services Manuf. J. 26 (4), 490–515.
Gkiotsalitis, K., Cats, O., 2020. Timetable recovery after disturbances in metro operations: An exact and efficient solution. IEEE Trans. Intell. Transport. Syst.
Goverde, R.M., 2010. A delay propagation algorithm for large-scale railway traffic networks. Transport. Res. Part C: Emerg. Technol. 18 (3), 269–287.
Gu, Y., Qian, Z., Chen, F., 2016. From Twitter to detector: Real-time traffic incident detection using social media data. Transport. Res. Part C: Emerg. Technol. 67 (3), 321–342.
Hansen, I.A., Goverde, R.M., van der Meer, D.J., 2010. Online train delay recognition and running time prediction. In: In 13th International IEEE Conference on Intelligent Transportation Systems. IEEE, pp. 1783–1788.
Hansen, I.A. (Ed.), 2008. Railway timetable & traffic: analysis, modelling, simulation. Eurailpress.
Harrod, S., Cerreto, F., Nielsen, O.A., 2019. A closed form railway line delay propagation model. Transport. Res. Part C: Emerg. Technol. 102, 189–209.
Huang, P., Wen, C., Fu, L., Lessan, J., Jiang, C., Peng, Q., Xu, X., 2020. Modeling train operation as sequences: A study of delay prediction with operation and weather data. Transport. Res. Part E: Logist. Transport. Rev. 141, 102022.
Huang, P., Li, Z., Wen, C., Lessan, J., Corman, F., Fu, L., 2021. Modeling train timetables as images: A cost-sensitive deep learning framework for delay propagation pattern recognition. Expert Syst. Appl. 177, 114996.
Jasperse, F.Z.G., 2020. Automated offline detection of disruptions using smart card data. TU Delft.
Jespersen-Groth, J. et al., 2009. Disruption Management in Passenger Railway Transportation. In: Ahuja, R.K., Möhring, R.H., Zaroliagis, C.D. (Eds.), Robust and Online Large-Scale Optimization. Lecture Notes in Computer Science, vol. 5868, Springer, Berlin, Heidelberg.
Ji, T., Fu, K., Self, N., Lu, C.T., et al., 2018. Multi-task learning for transit service disruption detection. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*. [Online] pp. 634–641.
Li, W., Peng, Q., Wen, C., Wang, P., Lessan, J., Xu, X., 2020. Joint optimization of delay-recovery and energy-saving in a metro system: A case study from China. Energy 202, 117699.
London Datastore, 2018. Data Quality Standards. Available from: https://data.london.gov.uk/about/data-quality-standards/ [accessed 14th September 2018].
Lord, E., Willems, M., Lapointe, F.J., Makarenkov, V., 2017. Using the stability of objects to determine the number of clusters in datasets. Inf. Sci. 393, 29–46.
Luo, S., He, S.Y., 2021. Using data mining to explore the spatial and temporal dynamics of perceptions of metro services in China: The case of Shenzhen. Environ. Plann. B: Urban Analy. City Sci. 48 (3), 449–466.
Mahmassani, H.S., Haas, C., Zhou, S., Peterman, J., 1999. *Evaluation of incident detection methodologies*. University of Texas at Austin. Centre for Transportation Research. Report number: FHWA/TX-00/1795-1.
Malandri, C., Fonzone, A., Cats, O., 2018. Recovery time and propagation effects of passenger transport disruptions. Phys. A: Statist. Mech. Appl. 505, 7–17.
Mattsson, L.-G., 2007. Railway capacity and train delay relationships. In: Murray, A.T., Grubesic, T.H. (Eds.), Advances in Spatial ScienceCritical Infrastructure. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 129–150.
McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. New York, N.Y. : M. Dekker.
Meester, L.E., Muns, S., 2007. Stochastic delay propagation in railway networks and phase-type distributions. Transport. Res. Part B: Methodol. 41 (2), 218–230.
Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N., 2013. A fuzzy Petri net model to estimate train delays. Simul. Model. Pract. Theory 33, 144–157.
Ni, M., He, Q., Gao, J., 2016. Forecasting the subway passenger flow under event occurrences with social media. IEEE Trans. Intell. Transp. Syst. 18 (6), 1623–1632.
Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., Anguita, D., 2018. Train delay prediction systems: a big data analytics perspective. Big Data Res. 11, 54–64.
Osorio-Arjona, J., Horak, J., Svoboda, R., García-Ruíz, Y., 2021. Social media semantic perceptions on Madrid Metro system: Using Twitter data to link complaints to space. Sustainable Cities Soc. 64, 102530.
Peel, D., McLachlan, G.J., 2000. Robust mixture modelling using the t distribution. Statist. Comput. 10 (4), 339–348.
Riveiro, M., Lebram, M., Elmer, M., 2017. Anomaly detection for road traffic: A visual analytics framework. *IEEE Trans. Intell. Transp. Syst.* 18 (8), 2260–2270.
Rossi, R., Gastaldi, M., Gecchele, G., Barbaro, V., 2015. Fuzzy logic-based incident detection system using loop detectors data. Transp. Res. Proc. 10, 266–275.
Santhosh, K.K., Dogra, D.P., Roy, P.P., 2020. Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey. *ACM Comput. Surv.* 53 (6), 1–26.
Sodemann, A.A., Ross, M.P., Borghetti, B.J., 2012. A review of anomaly detection in automated surveillance. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 42 (6), 1257–1272.
Spanninger, T., Trivella, A., Büchel, B., Corman, F., 2022. A review of train delay prediction approaches. J. Rail Transp. Plann. Manage. 22, 100312.
Steenbruggen, J., Tranos, E., Rietveld, P., 2016. Traffic incidents in motorways: An empirical proposal for incident detection using data from mobile phone operators. J. Transp. Geogr. 54, 81–90.
Sun, H., Wu, J., Wu, L., Yan, X., Gao, Z., 2016. Estimating the influence of common disruptions on urban rail transit networks. Transport. Res. Part A: Policy Pract. 94, 62–75.
Tessitore, M.L., Sama, M., D'Ariano, A., Hélouët, L., Pacciarelli, D., 2022. A simulation-optimization framework for traffic disturbance recovery in metro systems. Transport. Res. Part C: Emerg. Technol. 136, 103525.
Tonnelier, E., Baskiotis, N., Guigue, V., Gallinari, P., 2018. Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework. Neurocomputing. 298, 109–121.
Wei, X., Yang, Z., Liu, Y., Wei, D., Jia, L., Li, Y., 2019. Railway track fastener defect detection based on image processing and deep learning techniques: A comparative study. Eng. Appl. Artificial Intell. 80, 66–81.
Welankiwar, A., Sherekar, S., Bhagat, A.P., Khodke, P.A., 2018. Fault Detection in Railway Tracks Using Artificial Neural Networks. In: *Proceedings of the 2018 3rd IEEE International Conference on Research in Intelligent and Computing in Engineering, RICE 2018*. [Online]. 2018 IEEE. pp. 3–7.
Yaghini, M., Khoshraftar, M.M., Seyedabadi, M., 2013. Railway passenger train delay prediction via neural network model. J. Adv. Transport. 47 (3), 355–368.
Yang, X., Chen, A., Wu, J., Gao, Z., Tang, T., 2019. An energy-efficient rescheduling approach under delay perturbations for metro systems. *Transportmetrica B: Transport Dyn.* 7 (1), 386–400.
Yap, M., Cats, O., Krasemann, J.T., van Oort, N., Hoogendoorn, S., 2022. Quantification and control of disruption propagation in multi-level public transport networks. Int. J. Transp. Sci. Technol. 11 (1), 83–106.
Yu, J., Stettler, M.E.J., Angeloudis, P., Hu, S., Chen, X., 2020. Urban network-wide traffic speed estimation with massive ride-sourcing GPS traces. Transport. Res. Part C: Emerg. Technol. 112, 136–152.
Yuan, J., Hansen, I.A., 2007. Optimizing capacity utilization of stations by estimating knock-on train delays. Transport. Res. Part B: Methodol. 41 (2), 202–217.
Yuan, J., 2006. Stochastic Modeling of Train Delays and Delay Propagation in Stations, PhD dissertation, Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Transportation and Planning.
Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transport. Res. Part C: Emerg. Technol. 86, 580–596.
Zhang, X., Zheng, Y., Zhao, Z., Liu, Y., et al., 2021. Deep learning detection of anomalous patterns from bus trajectories for traffic insight analysis. Knowl.-Based Syst. 217, 106833.
Zhao, L., Wu, J., Ran, Y., 2012. Fault diagnosis for track circuit using AOK-TFRs and AGA. Control Eng. Pract. 20 (12), 1270–1280.
Zhou, P., Chen, L., Dai, X., Li, B., Chai, T., 2020. Intelligent prediction of train delay changes and propagation using RVFLNs with improved transfer learning and ensemble learning. IEEE Trans. Intell. Transp. Syst. 22 (12), 7432–7444.
Zhu, L., Guo, F., Krishnan, R., Polak, J.W., 2018. A Deep Learning Approach for Traffic Incident Detection in Urban Networks. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. [Online]. pp. 1011–1016.
Zulfiqar, O., Chang, Y., Chen, P., Fu, K., et al., 2020. RISECURE: Metro Security Incidents And Threat Detection Using Social Media. *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020*. [Online] pp. 531–535.