A data-driven approach for quantifying the resilience of railway networks

Knoester, Max J.; Bešinović, Nikola; Afghari, Amir Pooyan; Goverde, Rob M.P.; van Egmond, Jochen

DOI
10.1016/j.tra.2023.103913

Publication date
2023

Document Version
Final published version

Published in
Transportation Research Part A: Policy and Practice

Transportation Research Part A 179 (2024) 103913

Contents lists available at ScienceDirect

Transportation Research Part A

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

A data-driven approach for quantifying the resilience of railway networks

Max J. Knoester [a], Nikola Bešinović [a,*], Amir Pooyan Afghari [b], Rob M.P. Goverde [a], Jochen van Egmond [c]

[a] Department of Transport and Planning, Delft University of Technology, Delft, the Netherlands
[b] Safety and Security Science Section, Delft University of Technology, Delft, the Netherlands
[c] Traffic Management Division, ProRail, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Railways
Resilience
Bathtub model
Disruption management
Data-driven
ANOVA

ABSTRACT

Disruptions occur frequently in railway networks, requiring timetable adjustments, while causing serious delays and cancellations. However, little is known about the performance dynamics during disruptions nor the extent to which the resilience curve applies in practice. This paper presents a data-driven quantification approach for an ex-post assessment of the resilience of railway networks. Using historical traffic realization data in the Netherlands, resilience curves are reconstructed using a new composite indicator, and quantified for a large set of single disruptions. The values of the resilience metrics are compared across disruptions of different causes using Welch's ANOVA and the Games-Howell test. Additionally, representative resilience curves for each disruption cause are determined. Results show a significant heterogeneity in the shape of the resilience curves, even within disruptions of the same cause. The proposed approach represents a useful decision support tool for practitioners to assess disruptions dynamics and propose best measures to improve resilience.

1. Introduction

Railway networks enable large amounts of passenger and freight traffic every day. Under normal conditions, trains arrive and depart according to the timetable and only minor variations in the train service could commonly be observed. These minor variations are referred to as disturbances (Cacchiani et al., 2014). Larger variations involving an unexpected change due to the failure of infrastructure, breakdown of vehicles, unscheduled maintenance, extreme weather conditions or other external events are referred to as disruptions (Bešinović, 2020). While disturbances are handled by making adjustments only to the timetable, disruptions require additional adjustments to the rolling stock and crew planning (Mattsson & Jenelius, 2015; Zilko et al., 2016). The consequences of a disruption generally include train cancellations and significant delays. Due to the intrinsic characteristics of railway networks, disruptions can easily propagate through the network in time and space (Cats & Jenelius, 2014; Malandri et al., 2018) and their effects may even build up to a systemwide scale (Dekker & Panja, 2021). In that case, primary delays will have caused extensive secondary delays, and imbalances in the rolling stock and crew resources will have emerged, potentially leading to an out-of-control situation. The evolution of system performance during a disruption can be visualized in the 'resilience curve', which is illustrated schematically in Fig. 1. The resilience curve shows how performance first degrades and eventually recovers. Three phases are distinguished in this

* Corresponding author.
E-mail address: nikola.besinovic@tu-dresden.de (N. Bešinović).

https://doi.org/10.1016/j.tra.2023.103913
Received 25 May 2022; Received in revised form 12 October 2023; Accepted 18 November 2023
Available online 24 November 2023

curve: the first and the third phase (i.e. degradation and recovery, respectively) are transition phases, and the second phase (i.e. response) represents disrupted but stable system behavior. The resilience curve is sometimes referred to as the 'bathtub model' as well (Ghaemi et al., 2017).

Traffic control during disruptions is referred to as rescheduling and is commonly performed by the infrastructure manager. Rescheduling is anticipation-based in case recovery measures are predefined. However, tailor-made solutions have to be made for each disruption if rescheduling happens in real time. The latter puts more focus on the reactive capacity of train dispatchers and traffic controllers (Schipper & Gerrits, 2018). Adjusting the rolling stock and crew planning during a disruption is the responsibility of the train operating company (TOC). The joint actions taken by the infrastructure manager, TOCs and additional actors such as maintenance contractors and emergency services can be referred to as disruption management. For a detailed overview of the roles in the disruption management process, the tradeoffs in disruption management and the differences between countries, the interested reader may refer to Schipper and Gerrits (2018).

In the Netherlands, which has one of the busiest railway networks in Europe (ACM, 2019), traffic control is performed by the infrastructure manager, ProRail. Disruption management in ProRail is organized according to the bathtub model. In the first phase, emergency measures are taken regarding safety and logistics. In most cases, a contingency plan is applied which provides a revisited timetable for the second phase. In the second phase, the execution of the contingency plan is monitored and the cause of the disruption is resolved. Meanwhile, a recovery plan is prepared to resume the train service according to the original timetable. When the recovery plan is approved and the infrastructure is reclaimed, the recovery can be initiated. In the third phase, the execution of the recovery plan is monitored. This standardized process is followed regardless of the disruption cause. Because of the strong reliance on contingency plans, disruption management in the Netherlands is highly anticipation-based (Schipper & Gerrits, 2018). Nonetheless, disruption management remains a highly manual process in this country which requires extreme efforts by dispatchers and traffic controllers in real time, particularly in the first resilience phase.

Despite the fact that disruption management is organized according to the bathtub model, practical knowledge about system performance during disruptions is limited. In scientific literature as well, the resilience curve has mostly remained a theoretical concept, and thus, its interpretation has not resulted in major findings (Madni et al., 2020). The exact shape of the curve and the extent to which it applies in practice are not properly understood. Because of this limited quantitative knowledge, designing appropriate measures for disruption management has been a challenge (Bešinović, 2020). The complex interaction between delay propagation and management actions is most likely to blame for the relatively limited amount of past research (Büchel et al., 2020). However, there is currently a growing demand for the quantification of system performance during disruptions (Bešinović, 2020), as resilience has become a critical design requirement for increasingly complex and interconnected systems (Uday & Marais, 2015; Madni et al., 2020). Better quantitative knowledge on this topic would contribute to the effective allocation of resources to prevent, mitigate and recover from disruptions (Malandri et al., 2018). Therefore, the main research question in this study is: how does the system performance of a railway network develop during disruptions?

To answer this research question and address the gap, a data-driven quantification approach is presented in this study to make an ex-post assessment of the resilience of railway networks using historical traffic realization and disruption data. Several resilience metrics that quantify the shape of the resilience curve are extracted from the literature, such as the recovery time and performance loss, based on which two new resilience metrics are introduced: the degradation profile and the recovery profile. Using historical traffic realization data of railway operations in the Netherlands, the resilience curve is reconstructed and the phases in the resilience curve are identified for a large and heterogeneous set of frequently occurring disruptions in this country. Particularly, the five main disruption causes are considered, which are train defects, section/signal failures, collisions, switch failures and overhead line failures. Each resilience curve is described in terms of the resilience metrics. The values of the metrics are evaluated in statistical analyses in order to identify differences among disruptions of varying causes. The proposed approach is data-driven because it uses historical traffic realization data which most importantly specify the plan time and realization time of train activities. A train activity is defined here as an arrival, short stop, departure or passing of a train at a certain location in the network referred to as a timetable point. In comparison with other types of quantification approaches (e.g. topological, optimization-based, simulation-based), the benefits of this data-driven approach are: (1) it removes the need to model the traffic conditions in the network explicitly, and (2) it allows a direct comparison

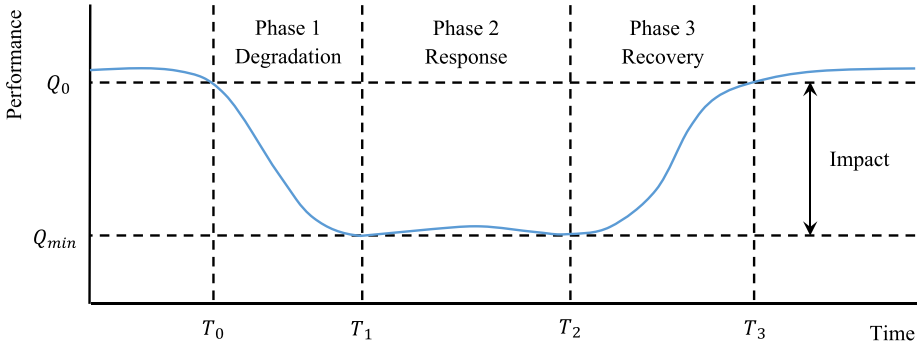


Fig. 1. Schematic illustration of the theoretical resilience curve.

between what practitioners believe to be true and what happens in reality.

This study contributes to the field of transport engineering and railway resilience from the following perspectives:

A new data-driven resilience quantification framework is developed to ex-post assess the resilience of railway networks.

A composite performance indicator is defined which is capable of representing delays and cancellations simultaneously.

A method is developed for identifying the resilience phases from the resilience curve.

A representative resilience curve is determined for the five most common disruption causes in the Netherlands through empirical testing on a real-life case study.

The remainder of this paper is organized as follows. Section 2 provides a literature review on the quantification of resilience in railways and other domains. Section 3 presents the proposed resilience quantification framework and discusses the methodology for the resilience quantification. Section 4 provides the case study description and presents the results of empirical testing of the proposed quantification approach. Section 5 provides a thorough discussion on the results, and Section 6 concludes the paper and presents future research directions.

2. Literature review

A reasonable number of studies have investigated the resilience of railway networks in the literature. A thorough review of these studies can be found in Bešinović (2020). In addition, Mattsson and Jenelius (2015), Zhou et al. (2019) and Gonçalves and Ribeiro (2020) have reviewed the resilience of transport systems in general. Based on these previous studies, a revisited definition of resilience is formulated as follows. Resilience of a railway system is the ability of the system to: (1) prepare for a disruption, as well as (2) reduce, absorb and accommodate the impact of a disruption while maintaining an acceptable level of service, and (3) recover to a desired state of operation within a reasonable amount of time. This definition fits the context of our study as it acknowledges the existence of different phases in a disruption, and also, it explicitly mentions the relevant levels of performance.

To better understand how resilience has been quantified in previous studies, we first review the existing resilience quantification approaches in Section 2.1. In general, different types of approaches include topological (e.g. Dorbritz, 2011; Lu, 2018; Xu and Chopra, 2022; Zhang and Ng, 2022), optimization-based (e.g. Van Aken et al., 2017; Bababeik et al., 2018; Bešinović et al., 2021; Bešinović et al., 2022; Liu et al., 2022), simulation-based (e.g. D'Lima and Medda, 2015, Adjetey-Bahun et al., 2016; Liu et al., 2021) and data-driven (e.g. Janić, 2018; Büchel et al., 2020; Yin et al., 2022). As described by Bešinović (2020), the majority of the research is optimization-based, while the number of studies of the other types of approaches is still limited. Here, we focus on data-driven approaches in railway transport and other transport modes including subway, taxi and air transport. In Section 2.2, we review the existing resilience metrics, which are used later for selecting appropriate metrics that describe the shape of the resilience curve. Section 2.3 concludes the literature review by highlighting the research gaps in railway resilience research.

2.1. Quantification approaches

With respect to railways, Chan and Schofer (2016) studied the recovery of the New York City metropolitan railway network in terms of revenue vehicle miles after several extreme weather events. Using the revenue vehicle miles, they defined the number of lost service days as an aggregate measure of resilience. Janić (2018) studied the recovery of the Japanese high-speed railway network after the 2011 earthquake by deriving an aggregate measure of resilience. This measure was composed of several performance indicators covering infrastructural, operational, economic and socio-economic aspects. Woodburn (2019) studied the consequences of a lengthy, unplanned closure of a major freight route in Britain by tracking the gradual improvement in traffic and service levels over a two-month period. Büchel et al. (2020) studied delay propagation in the Swiss railway network after a two-month disruption in Germany by comparing arrival delays for the disrupted and undisrupted scenario. The cascading effects over large distances were also replicated in a simulation. At a more abstract modelling level, Chen and Wang (2019) investigated the impacts of severe weather events on high speed rail and aviation delays; they did not explicitly consider cancelled services nor system performance evolution. Liu et al. (2021) developed a generic method for analyzing metro system risk based on historical incidents, which are used to generate low-probability, high-consequence scenarios. Yin et al. (2022) analysed historical realised traffic data and focused on train services. To evaluate resilience, they proposed a hybrid knowledge-based and data-driven Bayesian network.

With respect to other transport modes, Janić (2015) studied the resilience and friability of the air transport network around New York LaGuardia by deriving an aggregate measure of resilience, defined as the sum of the resilience of individual airports. Similarly, Janić (2019) proposed new analytical models reflecting interests of the main stakeholders for assessing the network resilience of airline cargo transport network affected by a large scale disruptive event. Zhu et al. (2016) studied the recovery of taxi and subway ridership in New York City after extreme weather events by assessing the loss of resilience per evacuation zone in terms of service capacity. Zhu et al. (2017) further investigated the spatial dependence of resilience per zone with a multivariate regression model. Ren et al. (2020) identified the relationships between causal factors and resilience with respect to disruptions in the Beijing subway network by constructing a Bayesian network. Wong et al. (2020) studied the resilience of individual airlines, rather than studying the network resilience, by searching for abnormalities in arrival delays in four US airlines. These abnormalities were quantified using a statistical measure called the Mahalanobis distance. Zhou and Chen (2020) measured airport resilience under various severe weather conditions and performed an empirical econometric analysis.

To summarize, previous data-driven studies on the resilience of transport systems mainly examined single, large-scale disruptive events which are relatively uncommon. Several existing research studied a large number of disruptions, but non did it to assess the evolution of system performance during disruptions.

2.2. Resilience metrics

An immediate question arising after plotting the resilience curve would be how to describe the evolution of performance, and thus, the shape of the resilience curve, quantitatively. Assuming that the exact mathematical function of the curve is not known, one way of doing so is to design a set of resilience metrics that are capable of summarizing how performance developed during the disruption. Since resilience is believed to be a multidimensional construct, it is incapable of being captured in a single metric (Munoz & Dunbar, 2015), and thus, multiple metrics are needed. If only one metric were to be used, then entirely different loss and recovery behaviors could result in the same resilience value (Zobel, 2011). Hosseini et al. (2016) made the distinction between deterministic and probabilistic resilience metrics. In this review, we only consider deterministic metrics.

Resilience metrics defined in a railway context include the recovery time (e.g. Chan & Schofer, 2016), recovery rate (Janić, 2018), deterioration rate (Janić, 2018), initial impact (Dorbritz, 2011), maximum impact (e.g. Nicholson et al., 2015) and minimum performance (Dorbritz, 2011). Minimum performance may also be referred to as residual functionality (Cimellaro et al., 2010). Recovery time is the most common metric in transport literature (Zhou et al., 2019). Another common metric is the area above the resilience curve, which is known by different names such as the deviation area (Nicholson et al., 2015), service loss (Chan & Schofer, 2016), or originally, loss of resilience (Bruneau et al., 2003). Resilience metrics are found to be domain-independent, which explains why the discussed metrics also appear in studies that investigate the resilience of systems in general.

For a broader perspective, resilience metrics in supply chain literature are explored, which introduces additional metrics not encountered in transport literature. Spiegler et al. (2012) adopted the integral of time absolute error (ITAE) commonly applied in control engineering. Munoz and Dunbar (2015) defined the profile length and the weighted sum to describe the nonlinearity of the resilience curve, although in their interpretation, the drop in performance is abrupt and performance recovers gradually. In that case, the resilience curve could be perceived as consisting only of a third phase. The weighted sum is defined in Munoz and Dunbar (2015) as the time-dependent deviation from a linear recovery. An overview of the resilience metrics used in previous studies, including those used in general systems and supply chain research, is presented in Table 1.

2.3. Research gaps

Previous data-driven studies on the resilience of railway networks have left a number of research gaps. First, the evolution of railway system performance during the consecutive resilience phases is not well understood for disruptions of varying scale and origin. The reviewed studies addressed mostly single, large-scale disruptions such as earthquakes and storms. In reality though, disruptions of a smaller scale such as switch failures or train defects occur frequently, and yet, these disruptions have not been subject to resilience research.

Second, because no earlier attempts have been made to reconstruct the resilience curve for a large number of disruptions, no effort has been made to develop an efficient and accurate method for identifying the resilience phases from the curve. Such a method, which would produce the start and end time of each resilience phase, would be required to determine the values of phase-specific resilience metrics.

Third, realization data have not been used to assess the resilience of a railway network for a large and heterogeneous set of disruptions. While modern technologies and data analytics create opportunities for the use of empirical data (Parkinson & Bamford, 2017), only Ren et al. (2020) collected data for a large and heterogeneous set of disruptions in a railway-like context. The lack of useful reference material in this area poses challenges with regard to data collection, preparation and analysis.

Fourth, the spatial attributes of a railway network have not been addressed explicitly when studying resilience as a function of time. This is demonstrated by the fact that the dimension of time is usually presented on a separate axis, whereas the dimension of space is not considered explicitly. The reviewed studies provide insufficient insights into how the spatial attributes of a network can be properly accounted for in the calculation of railway system performance.

Table 1
Resilience metrics used in previous studies.

Resilience metric	Research domain	References
Recovery time	General systems, railways, supply chain	Chan & Schofer (2016), Dorbritz (2011), Janić (2018), Munoz & Dunbar (2015), Nicholson et al. (2015), Ouyang et al. (2012), Zhou et al. (2019), Zobel (2011)
Recovery rate	General systems, railways	Cimellaro et al. (2010), Janić (2018)
Deterioration rate	Railways	Janić (2018)
Initial impact	General systems, railways	Dorbritz (2011), Ouyang et al. (2012), Zobel (2011)
Maximum impact	General systems, railways, supply chain	Janić (2018), Munoz & Dunbar (2015), Nicholson et al. (2015), Ouyang et al. (2012)
Residual functionality	General systems, railways	Cimellaro et al. (2010), Dorbritz (2011)
Performance loss	General systems, railways, supply chain	Bruneau et al. (2003), Chan & Schofer (2016), Munoz & Dunbar (2015), Nicholson et al. (2015), Zhu et al. (2016)
ITAE	Supply chain	Spiegler et al. (2012)
Profile length	Supply chain	Munoz & Dunbar (2015)
Weighted sum	Supply chain	Munoz & Dunbar (2015)

3. Methodology

A data-driven methodology is developed in this study to ex-post evaluate the resilience of railway networks given a large number of disruptions of varying types. The methodology is generic and can work for different performance indicators, resilience metrics and categorizations of disruption types. It is structured in the resilience quantification framework presented in Fig. 2. The framework has been divided into three parts (i.e. input, processing and output) and works as follows. The quantification starts with the collection of traffic realization data, disruption log data and network data from a database. Traffic realization data includes at least the plan time, realization time and location of train activities. Disruption log data refer to information about the disruption, such as the reported start and end time, location, cause, etc. Network data specify how each timetable point in the network is connected to neighboring timetable points. Using these data, the evolution of performance over time can be calculated for each disruption in a disruption-specific impact area. Performance measurements at each time instant are stored in a dataframe, which serves as the basis for calculating the resilience metrics. The metrics are stored in a separate dataframe, which serves as the basis for conducting statistical analyses. The statistical analyses are used to identify differences among disruptions of varying types. The resilience metrics and the test statistics represent the numerical output of the resilience quantification. The resilience curves, which can be drawn directly from the performance measurements, represent the graphical output of the resilience quantification. Apart from plotting individual resilience curves, one could also identify different shapes of resilience curves and calculate the mean and median resilience curve per disruption type.

The process blocks in the resilience quantification framework are explained in more detail in the remainder of this section. Section 3.1 explains how the impact area is determined. Section 3.2 explains how performance is defined and how it is calculated. Section 3.3 explains which resilience metrics are selected and how they are defined. Section 3.4 describes the statistical methods applied in the statistical analyses.

3.1. Determining the impact area

A question arising before calculating the performance during a disruption is which area to consider in this calculation. Since disruption effects can easily spread through the network, it is preferred to study a larger area than just the disrupted line or timetable point. If the studied area is too small, then disruption effects further away from the disruption location could be overlooked, but if the studied area is too large, the impact of the disruption becomes less visible. Disruption management in the Netherlands provides a theoretical foundation for determining which area to consider, based on the concepts of decoupling points and impact areas. A decoupling point is defined as a timetable point where trains are allowed to start or end their route in case of a disruption. We first introduce the boundaries of a disruption, which could be the decoupling points closest to the disruption location, but could also be the timetable points marking a smaller area in which there is no traffic operation, depending on the magnitude of the disruption. The first impact area is bounded by the first intercity decoupling points from the disruption location; the second impact area is bounded by the

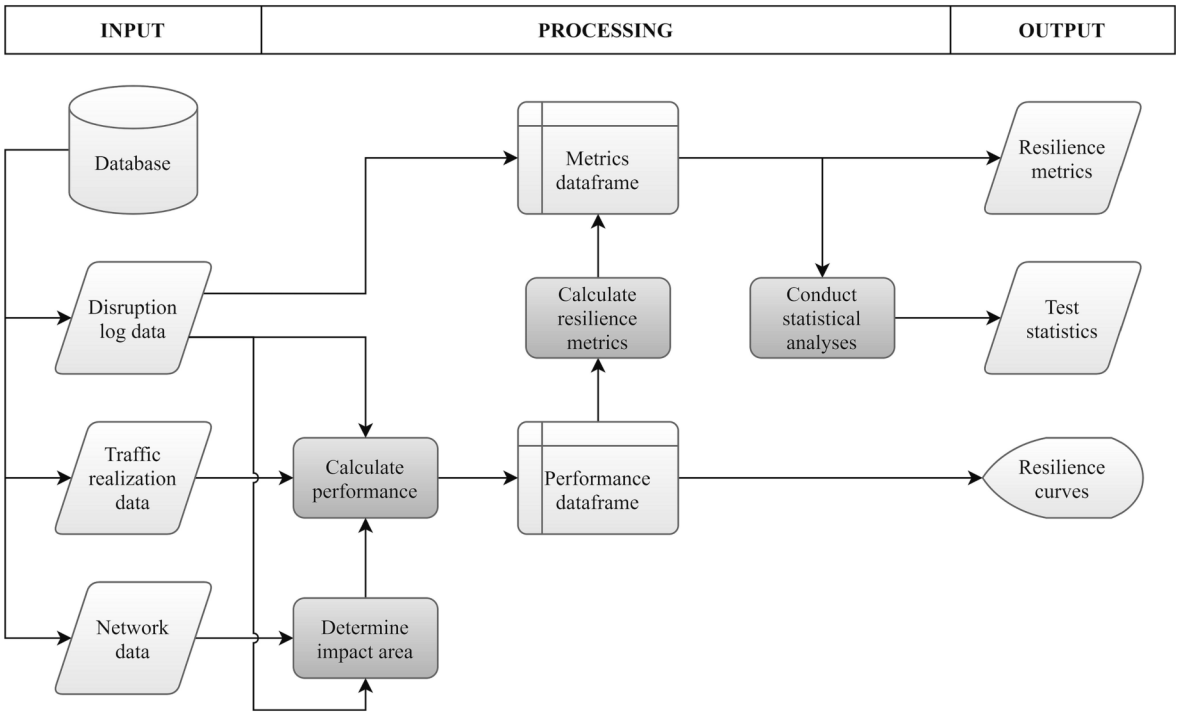


Fig. 2. The resilience quantification framework.

next closest intercity decoupling points from the first ones; and the third impact area is bounded by the next closest intercity decoupling points from the second ones. Since cancellations in the third impact area are in principle not allowed, disruption effects are mostly contained within the first and second impact area. Thus, the first and second impact area are identified as the appropriate area to consider for the resilience quantification. An additional benefit of using impact areas is that it limits possible interactions with disruptions in other parts of the network, both in our quantification approach and in practice. The size of the area depends on the location in the network, but also on the type of impact, as illustrated in Fig. 3.

To determine the impact area of each disruption given the disruption location, a modified breadth first search (BFS) algorithm is developed. The original BFS algorithm, introduced by Moore (1959), is a type of graph search that is used to traverse a graph and find each node or vertex in the graph. The BFS algorithm starts from a source node, referred to as the start vertex, and visits any adjacent, unvisited vertices until none are left. For our railway network context, several new extensions are required on the basic BFS algorithm. In the modified BFS algorithm, the start vertex/vertices is/are the disruption location (in case of a timetable point outage) or boundaries (in case of a line blockage), and the end vertices are the second closest decoupling points. The area covered in between represents the impact area. The new extensions are set as follows. First, the algorithm terminates its search along a branch of the network when a second decoupling point is reached in the specific branch. Second, the algorithm cannot follow a path that is not driven by any train. Third, the algorithm must account for the fact that not all vertices may appear in the same path in opposite directions. Fourth, the algorithm must allow a vertex to have more than one parent (i.e. the previous vertex in the specific path), and to visit this vertex when it has already been visited from another parent. Fifth, the algorithm needs to handle more than one start vertex, since a single start vertex only occurs for a timetable point outage. Instead, a line blockage will have two, three or even four start vertices, depending on the network layout. For example, in case of a line blockage between Amersfoort (Amf) and Apeldoorn (Apd) as in Fig. 3 (a), line Amf-Apd represents a simple, open line, and thus, two start vertices would suffice (being Amf and Apd) which would also be marked as the first decoupling points. Alternatively, assuming a disrupted station in Apd as in Fig. 3 (b), one start vertex would suffice (being Apd). If no train operations would be possible to and from Apd, this would affect the surrounding lines significantly, and the first decoupling points would be Amf, Dv and Zp. Instead, if limited train operations to and from Apd would be possible, then Apd itself would be the first decoupling point, and Amf, Dv and Zp would be the second decoupling points. The formal procedure of the modified BFS algorithm for a single start vertex is included in Appendix A. For more details on the algorithm, the interested reader may refer to Knoester (2021).

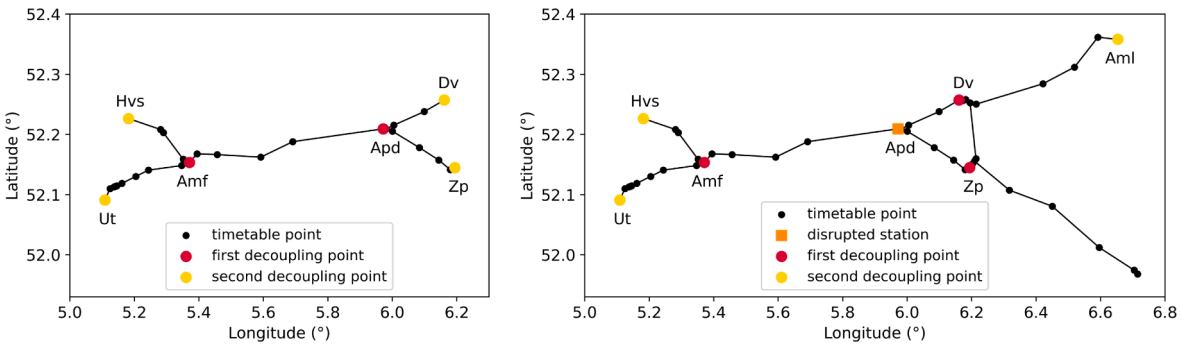
3.2. Calculating performance

In order to calculate performance, it is necessary to specify a performance indicator which measures the level of operations at a given time instant. A common approach is to express railway system performance in terms of the traffic level (e.g. Ghaemi et al., 2017), which we refer to as the traffic intensity. Traffic intensity is defined as the proportion of realized train activities relative to the scheduled train activities in a given time period. While such a definition for the traffic intensity indicator accounts for cancellations, it does not account for delays. Therefore, we also consider the traffic punctuality, which is defined as the proportion of punctual train activities (i.e. with a delay less than three minutes, which is a threshold used at the Dutch railways) relative to the realized train activities in a given time period. Together, the performance indicators account for cancellations as well as delays.

To measure both indicators simultaneously, they are combined in a composite performance indicator Q which is calculated as the weighted sum of traffic punctuality (P/R) and traffic intensity (R/S):

$$Q = \left((1-\lambda)\frac{P}{R} + \lambda\frac{R}{S} \right) \bullet 100[\%], \tag{1}$$

where P is the number of punctual train activities, R is the number of realized train activities, S is the number of scheduled train



(a) Partial/full line blockage. (b) Full timetable point outage.

Fig. 3. Timetable points in the first and second impact area for (a) a line blockage between Amersfoort (Amf) and Apeldoorn (Apd), and (b) a full timetable point outage in Apeldoorn (Apd). Note: there is no second decoupling point along the line from Apd to Zp, because the last node marks the end of the line in Winterswijk, and as such it is not classified as a decoupling point by ProRail.

activities and λ is the normalized performance weight. Thus, the punctuality is measured by the number of punctual activities relative to the number of realized activities. A weight $0 < \lambda < 0.5$ puts more emphasis on punctuality, whereas a weight $0.5 < \lambda < 1$ puts more emphasis on traffic intensity. Although the composite indicator is somewhat abstract and less easy to communicate than a single indicator, it has specific benefits. First, it makes it possible to measure delays and cancellations simultaneously. Second, it can smoothen fluctuations in these indicators, which are most prominent in the punctuality component. Third, the composite indicator helps account for the fact that not every disruption has the same impact on the train service: some disruptions may be more impactful in terms of delays, while others may be more impactful in terms of cancellations. Fourth, the composite indicator helps identify the start of a disruption more accurately in case measures are not taken immediately and delays start to build up as a result. If only the traffic intensity were to be studied, these effects would not be observed.

Performance at each time instant is calculated according to Eq. (1) as a centered moving average over a time period of 30 min and with a step size of one minute. These values are taken after initial testing of the approach. A smaller time period would be more capable of showing the dynamics in system performance, but would also make the resilience curve more difficult to analyze. The proposed indicator Q holds for dense railway networks/areas with many trains, i.e. $S \gg 10$ train activities, as it is the case in e.g. The Netherlands. Performance calculation using Eq. (1) is performed on a subset of the realization data containing only those train activities in the specified impact area.

3.3. Calculating the resilience metrics

The shape of the resilience curve is described by seven resilience metrics. The degradation time (*DT*), response time (*RST*) and recovery time (*RCT*) describe the duration of the first, second and third resilience phase, respectively. The maximum impact (*MI*) describes the vertical distance between target performance (Q_0) and minimum performance (Q_{min}). Performance loss (*PL*) describes the area enclosed by target performance and the resilience curve. Finally, the degradation profile (*DP*) and recovery profile (*RP*), which are based on the weighted sum in Munoz and Dunbar (2015) but do not include a time penalty, describe the summed deviation from a linear degradation and recovery, respectively. The seven resilience metrics are defined as follows, using the notation indicated in Fig. 1:

$$DT = T_1 - T_0, \tag{2}$$

$$RST = T_2 - T_1, \tag{3}$$

$$RCT = T_3 - T_2, \tag{4}$$

$$MI = Q_0 - Q_{min}, \tag{5}$$

$$PL = \sum_{0 \le i \le n-1 \; Q(t_i) < Q_0} (Q_0 - Q(t_i))(t_{i+1} - t_i), T_0 \le t_i \le T_3, \tag{6}$$

$$DP = \sum_{j=0}^{m} \left(f(t_j) - Q(t_j)\right), T_0 \le t_j \le T_1, \tag{7}$$

$$RP = \sum_{k=0}^{s} (g(t_k) - Q(t_k)), T_2 \le t_k \le T_3, \tag{8}$$

where T_0 is the start time of the disruption; T_1, T_2 and T_3 are the end time of the first, second and third phase, respectively; $f(t)$ is the linear degradation function connecting (T_0, Q_0) and (T_1, Q_{min}); $g(t)$ is the linear recovery function connecting (T_2, Q_{min}) and (T_3, Q_0); n is the total number of time intervals $t_{i+1} - t_i$, which is equal to the disruption length when taking a step size of one unit; and m and s are the number of equally spaced measurement points in the first and the third phase, respectively. The first five metrics can only have nonnegative values, whereas the last two metrics can have positive values as well as negative values. For all metrics, a higher positive value indicates a stronger disruptive effect, and thus, a less resilient network. Looking at the degradation profile (*DP*), in case that it degrades gradually and takes a long time to reach minimum performance, the *DP* value will be smaller than when performance drops quickly and then keeps degrading slowly. Therefore, such system is deemed more resilient, because the gradual degradation gives traffic controllers more time to respond, hence the smaller *DP*. Similar holds for recovery profile (*RP*).

The current DP and RP metrics work well under the assumption of convex/concave behaviour of the performance in degradation/recovery phases, which is the most commonly observed behaviour in real-life railway disruptions. However, due to a possible complex performance variability during some specific disruptions (e.g. a sudden drop in performance during recovery phase), DP or RP could reach close to zero values. Hence, it is recommended to always draw the resilience curve when studying a particular disruption, before interpreting its resilience metric values."

In order to calculate the resilience metrics, it is necessary to identify the timepoints $T_0, \cdots, T_3$ which are shown in Fig. 1. All timepoints are derived from the resilience curve, because the reported timepoints in the disruption log data may not always be accurate. Here, the start of disruption T_0 is defined as the last moment before the reported start of the disruption when performance is still above target Q_0. The end of disruption T_3 is defined as the first moment after the recovery is initiated when performance is above target

Q_0 again. Target performance Q_0 is defined as the average networkwide performance during the day, measured over a number of relatively quiet days (i.e. with a limited number of disruptions) throughout the year.

Assuming an approximately steady second phase, its start and end time T_1 and T_2 are determined by a modified steady state detection algorithm based on the approach of Dalheim and Steen (2020). A steady state detection algorithm is generally used to identify the steady parts of time series data. The approach of Dalheim and Steen (2020) involves performing regression analysis on consecutive, overlapping time windows. Their approach is modified to account for the dynamic nature of railway system performance, which results from the interplay between rescheduling the timetable, rolling stock and crew, and also for example from the up- or downscaling of a contingency plan and from deviations from the plan. Our modified steady state detection algorithm works by performing regression analysis on consecutive, overlapping time windows in the lower part of the resilience curve. Modifications to the approach of Dalheim and Steen (2020) include the following. First, the length of the time windows is made dependent on the total number of measurement points by taking the square root of this number and multiplying it by a scaling factor δ. Taking the square root allows short windows for relatively short disruptions and longer windows for longer disruptions, while it prevents too long windows for very long disruptions. The scaling factor δ is set to 1.8 based on initial testing of the algorithm, after establishing that lower values (e.g. $\delta = 1.5$) can yield too short windows, and higher values (e.g. $\delta = 2.0$) can yield too long windows occasionally. Second, the identification of a steady state for a particular window is only allowed in case the first performance measurement in the window is below a certain threshold, to avoid a steady state from being detected in the first or third resilience phase. Third, evaluation by the sliding window from Dalheim and Steen (2020) is dropped, since the algorithm only has to produce the first and last steady point in the lower part of the resilience curve. The formal procedure of the modified steady state detection algorithm is included in Appendix B. For more details on the algorithm, the interested reader may refer to Knoester (2021).

3.4. Conducting statistical analyses

Statistical analysis is a means to investigate patterns and relationships in quantitative data. Group comparisons are a class of statistical analysis, which we use to identify differences in the resilience metrics among disruptions of varying types. Knowing where and how large these differences are could help improve resilience in specific types of disruptions and in a specific phase or phases of a disruption. First, it is necessary to categorize the disruptions into groups. The disruption cause is taken as the variable that defines group membership, since for example, a line blockage due to a train defect might be inherently different than a line blockage due to an overhead line failure.

The standard parametric option for group comparisons of a single dependent variable is one-way analysis of variance (ANOVA), while the nonparametric alternative is the Kruskal-Wallis test. In both cases, a number of assumptions must be satisfied in order to draw justified conclusions from the test results. Since the distributions of the resilience metrics violate the assumptions of both one-way ANOVA and the Kruskal-Wallis test, we use Welch's ANOVA for the group comparisons. This method is similar to one-way ANOVA, but applies weights to adjust the grand mean (i.e. the mean of the total sample) based on the group means. Since ANOVA is robust against violation of the normality assumption (Mertens et al., 2017), Welch's ANOVA is useful for analyzing data that are nonnormally distributed and that have unequal variances among groups. The F-statistic, which describes the part of the variation in the dependent variable that is explained by group membership, is defined as follows:

$$F = \frac{SS/(g-1)}{1 + \frac{2\Lambda(g-2)}{3}}, \tag{9}$$

where SS is the weighted sum of squares of the residual error, g is the number of groups and Λ is a factor based on the weights and group sizes. A high F-statistic indicates that differences among groups likely exist. ANOVA is an omnibus test, which is two-sided by definition. This means it reveals whether a difference exists among groups, but not where exactly the difference lies or how large it is (Mertens et al., 2017). A post hoc test is required to explore the results in more detail. The common post hoc test for Welch's ANOVA is the Games-Howell test, which applies a series of pairwise comparisons among the groups while controlling for the family error rate. This test is similar to Tukey's Honest Significant Difference, but does not require equal variances among groups.

4. Case study and results

The proposed methodology is applied to a case study on disruptions in the Dutch railway network. The case study consists of several experiments designed to incrementally gain a better understanding of the performance dynamics. First, Section 4.1 defines the scope of the case study. Section 4.2 presents the detailed evaluation of an exemplary disruption to illustrate the working of the resilience quantification framework. Section 4.3 presents the different shapes of resilience curves that are recognized. Section 4.4 presents the mean and median resilience curve per disruption cause. Section 4.5 presents the results of the group comparisons of the resilience metrics.

4.1. Case description

The case study focuses on passenger traffic in the Netherlands in timetable year 2019, which was the last regular timetable year before the COVID-19 pandemic. Passenger traffic includes regional, intercity, high-speed and international trains. The study area includes the entire Dutch railway network with the exception of traffic control area Kijfhoek that exclusively handles freight traffic. In

total, 2,152 disruptions occurred in the study area and time period for which a capacity reallocation and usually also a contingency plan were applied. The five most common disruption causes are studied, which together accounted for 76 % of the disruptions in 2019. In descending order of occurrence, these causes include train defects, section/signal failures, collisions, switch failures and overhead line failures. Section failures and signal failures are regarded as a single category because they are often related. Also, collisions are regarded as one general category, including collisions with a person, (motor) cyclist, road vehicle, animal and infrastructure object. Collisions with a person are by far the most common type of collision. Regarding the train activities, only the arrivals, short stops and passings are included, since including the departures as well would mean that a train is observed twice at the same location when it makes a stop. After filtering the realization data for the specified activities, the data contain approximately 132,000 activities per day.

To prevent the influence of other disruptions in the resilience curve, we only consider single disruptions, which are defined as disruptions that do not have overlapping time periods nor common timetable points in their impact areas. The single disruptions that are studied usually start and end on the same day, occur on normal days of operation, and can be compared based on the time estimates currently used in practice. In addition, the studied disruptions have a resilience curve which drops below target performance Q_0 at some point during the disruption, and they have an observed start and end time within a reasonable proximity to the reported start and end time, respectively, to reduce a potential influence of other disruptions on the considered resilience curve.

Given the focus of our study, the exemption criteria are as follows. Disruptions are excluded from the resilience quantification if they: do not match the top five causes; occurred on a black day (i.e. a day when traffic punctuality is below 75 % and/or traffic intensity is below 90 % networkwide) or a near-black day; have overlapping time periods and impact areas; have a negligibly small impact area (due to the absence of passenger traffic); have a limited impact (i.e. performance did not drop below Q_0); have a reported duration longer than ten hours; have a missing time entry for the start of recovery; have a start time which cannot be determined within 60 min before the reported start of disruption (e.g. due to an earlier disruption or disturbance); have an end time which cannot be determined within 180 min after the reported end of disruption (e.g. due to a follow-up disruption or disturbance); have zero scheduled activities in any given time window (due to the disruption starting or ending early morning or late night); or have no identifiable steady state.

Table 2 shows an overview of the number of all disruptions, single disruptions and single disruptions analyzed in this study per disruption cause. Note that the number of excluded disruptions based on the exemption criteria partly depends on how performance evolved, and thus, on the value of the performance weight λ. In some cases, it could be that the target performance is not met at start and/or end of disruption or that a single disruption turned out to be a connected disruption because of a longer observed duration. Therefore, we tested the selection of disruptions for various values of λ and obtained comparable results. Here we report the number of disruptions studied taking into account $\lambda = 0.67$. The choice of this value of λ, which puts twice the weight on the traffic intensity compared to punctuality, is elaborated on in Section 4.2. Finally, target performance Q_0 is set to 97 % after observing performance during fairly normal traffic conditions. We initially investigated a possible existence of hourly variations in Q_0, however such was not observed. Therefore, the selected Q_0 represents an average of daily network performance.

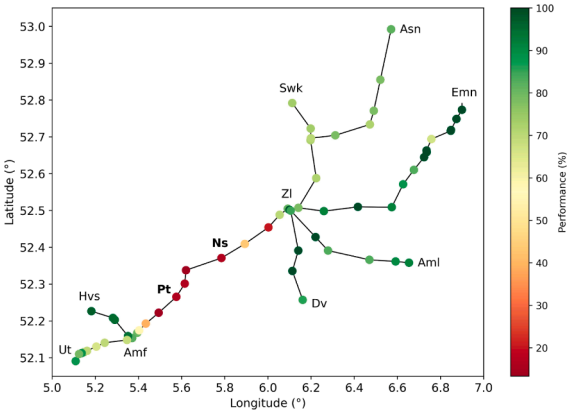
4.2. Characteristics of the resilience curve for an example case

An exemplary disruption is selected to demonstrate the working of the resilience quantification framework up until statistical analysis. In particular, we present: (1) the effects of the performance weight λ on the resilience curve, (2) the performance dynamics during disruption, (3) the outcome of our modified steady state detection algorithm, and (4) the values of the resilience metrics. The example case is a collision that occurred between Putten and Nunspeet on April 4, 2019. The disruption location and the impact area are shown in the network diagram in Fig. 4 (a). The collision reportedly occurred at 14:43 and resulted in a full line blockage of the double track line. At 16:24, the stranded passengers were evacuated to Amersfoort by a replacement train. At 16:56, the damaged train was brought to Amersfoort after being cleaned on site, and later that night it was brought to Amsterdam for extensive cleaning. The first impact area (bounded by Amersfoort and Zwolle) and the second impact area (bounded by Steenwijk, Assen, Almelo, Deventer, Hilversum and Utrecht) comprised an impact area of 64 timetable points. During the 223-minute disruption, there were 2,930 scheduled arrivals, short stops and passings in this area, of which 2,369 (80.85 %) were realized. Of the realized activities, 2,206 (93.12 %) were punctual.

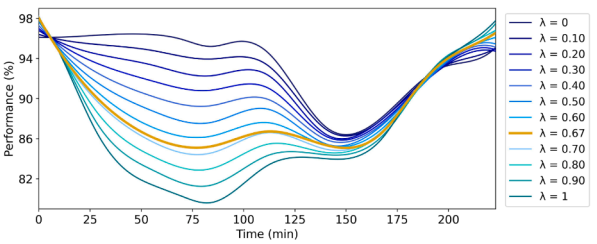
Performance is calculated firstly for different values of the performance weight. The resulting resilience curves are presented in Fig. 4 (b) as cubic splines fitted to the actual performance measurements. The two extremes show performance measured based on punctuality only ($\lambda = 0$) and traffic intensity only ($\lambda = 1$). It is observed that traffic intensity dropped quickly at the start of the disruption, while punctuality remained relatively stable until approximately 100 min into the disruption. The moment when punctuality eventually dropped matches the moment when evacuation of the stranded passengers began. Meanwhile, traffic intensity had

Table 2
Number of disruptions matching the five most common causes.

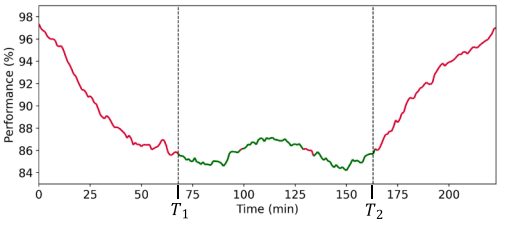
Disruption cause	No. of disruptions in the dataset	No. of single disruptions	No. of single disruptions studied ($\lambda = 0.67$)
Train defect	742	346	202
Section/signal failure	306	141	97
Collision	275	146	96
Switch failure	153	47	34
Overhead line failure	65	26	16
Total	1,541	706	445



(a) Network diagram showing the boundaries of the disrupted area (printed in bold), the first and second impact area, and the minimum performance level at each timetable point for .



(b) Resilience curve for different values of the performance weight.



(c) Resilience curve for $\lambda = 0.67$ with the identified steady state.

Fig. 4. Visualizations for the exemplary disruption, showing (a) minimum performance in the first and second impact area, (b) the evolution of performance for different values of the performance weight, and (c) the outcome of the steady state detection algorithm.

already partly recovered. Performance recovered similarly in the third phase for the different values of the performance weight, which indicates that the recovery was managed well and did not cause many new delays. Much lower weights than $\lambda = 0.67$ would underestimate the impact in terms of traffic intensity, while much higher weights would neglect the good performance in terms of punctuality. This provides additional support for our choice of λ, and thus, $\lambda = 0.67$ is maintained throughout the experiments. Finally, this value of λ was supported by the planners at Dutch infrastructure manager ProRail, where it is more important that trains are running, albeit with certain delays. Other companies could use a different weighing factor to balance the two components differently.

Fig. 4 (c) shows the outcome of the modified steady state detection algorithm on the resilience curve for $\lambda = 0.67$ (the yellow line in Fig. 4 (b)), resulting in the timepoints T_1 and T_2. The steady parts of the curve are shown in green, while the unsteady parts are shown in red. The detection of a steady state is successful, as it matches the steady state that one would identify by observation, and also, it is not affected by the slight change in performance during the second phase. Note that in practice, it is nearly impossible to observe the

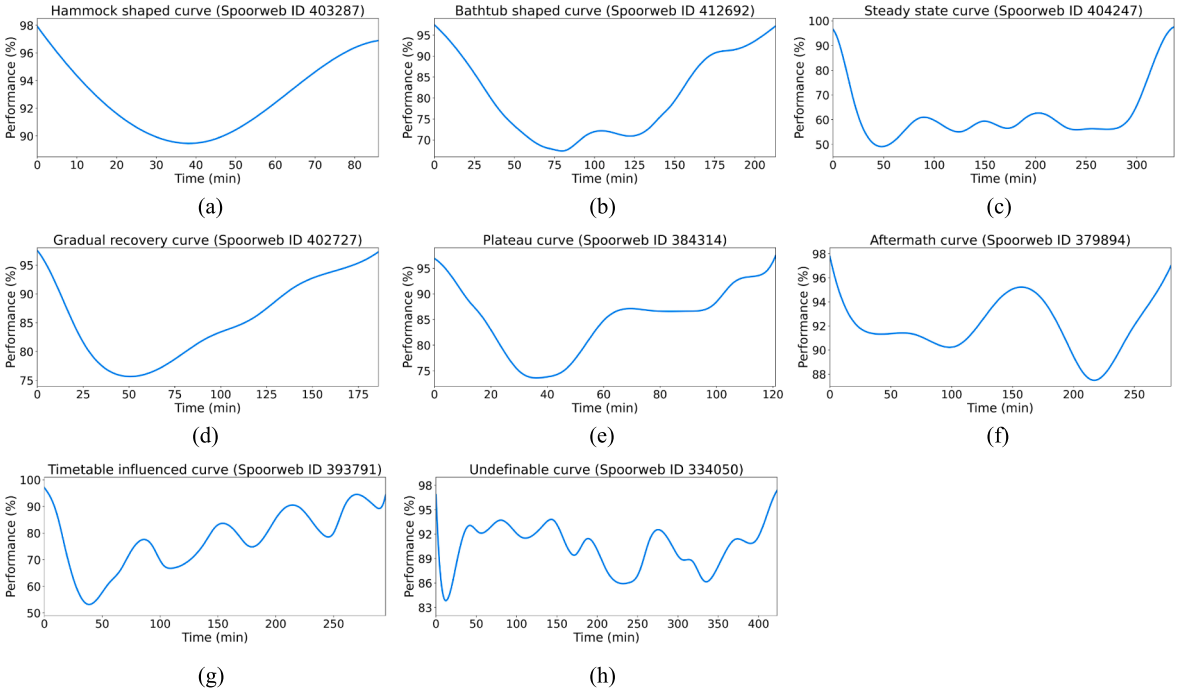


Fig. 5. Examples of the different shapes of resilience curves.

response phase as a predominantly straight line, as in the theoretical representation in Fig. 1. Fig. 4 (c) represents a "raw" resilience curve, without smoothening, and Fig. 4 (b) and 5 have been smoothed using cubic splines.

Based on the resilience curve in Fig. 4 (c), the resilience metrics are calculated. The degradation time, response time and recovery time equal 68, 95 and 60 min, respectively. The maximum impact equals 12.80 % and the performance loss equals 1,848 min. The degradation profile equals 108.53 %, indicating a convex deviation from a linear degradation. This means performance dropped rapidly due to the cancellation of trains early in the disruption. The recovery profile equals −47.44 %, indicating a smaller, concave deviation from a linear recovery. This means performance recovered rapidly as many trains could be reinserted shortly after the recovery was initiated.

4.3. Different shapes of resilience curves

Although the resilience curve is depicted in theory as a bathtub shaped curve with a clearly recognizable first, second and third phase, other shapes are possible in practice as well. Inspection of over 100 randomly sampled disruptions (at least ten per disruption cause) reveals that it is possible to distinguish between eight shapes of resilience curves, including the bathtub. Real examples of the different shapes of curves are presented in Fig. 5, which are portrayed as cubic splines fitted to the actual performance measurements. We recognize the following shapes:

a. The hammock shaped curve, which follows a smooth transition from degradation to recovery without a distinctive, steady second phase.
b. The bathtub shaped curve, which is similar in appearance to the common depiction of the resilience curve in theory.
c. The steady state curve, which shows a dominating and distinctive, steady second phase.
d. The gradual recovery curve, which shows a gradual recovery that starts directly after the first phase without a distinctive, steady second phase.
e. The plateau curve, which recovers well initially but takes a long time to fully recover to target performance.
f. The aftermath curve, which recovers well initially but shows a drop in performance towards the end. Afterwards, the curve quickly recovers to target performance.
g. The timetable influenced curve, which may resemble any of the other types of curves but also shows a periodic variation introduced by the timetable.
h. The undefinable curve, which represents the cases that seem to defy all logic and do not fit any of the previous descriptions.

The hammock shaped curve, bathtub shaped curve and steady state curve resemble the theoretical representation of the resilience curve in Fig. 1 the most, and could therefore be referred to as the short, moderate and long bathtub, respectively. The hammock shaped curve could be observed for relatively short disruptions, whereas the steady state curve could be observed for particularly long disruptions, in which case it is clear that there will be limited traffic for an extended period of time.

The shape of the gradual recovery curve may be explained by the fact that many trains tend to be canceled shortly after a disruption is reported. While doing so, the number of initial cancellations could be unnecessarily high, which could lead to trains being gradually reinserted right after, creating a distorted version of the bathtub. This shape of resilience curve appears most typical for collisions, and is also occasionally observed for train defects. This observation contradicts the belief among traffic controllers and analysts that collisions always have a dominating and distinctive second phase due to the necessary clearing operations.

The shape of the plateau curve may be explained by the application of a new, less restrictive contingency plan, or by deviating from the existing plan in anticipation of the recovery. These actions could increase performance during the disruption to a level that is higher than minimum performance, but still below target.

The aftermath curve includes a significant drop in performance towards the end. This shape could be observed for infrastructure related disruptions, and for switch failures in particular, where the drop appears to be related to the additional closure of tracks for the permanent repair of the infrastructure. In this case, a higher number of trains needs to be canceled temporarily to create a safe and accessible workspace for the mechanics team.

The timetable influenced curve is characterized by an hourly pattern which is imposed on the resilience curve. In the specific example in Fig. 5 (g), the underlying curve is a gradual recovery curve. Since the underlying curve can have different shapes, the timetable influenced curve may be considered not as a distinctive shape itself, but rather as a result of the number of train activities in the observed time period. This type of curve could occur for disruptions that involve relatively little traffic because they occur in a more remote part of the network and/or occur in the early morning or late evening, when train frequencies are low. Note that the wave pattern in some of the curves (e.g. Fig. 5 (g)) occurs due to the varying number of train activities in subsequent time intervals. In particular, at every timestep, performance is calculated for an interval of 30 min. As time progresses, the number of activities in the time interval changes. Depending on the frequency of trains (e.g. trains with low frequency running every 60′) and the volume of train traffic in the studied area (i.e. low volumes), this could introduce a wave pattern in the resilience curve.

The undefinable curve can be an outcome of various uncertainties during the disruption. For example, it may be explained by an unclear nature and/or location of the failure, which can result in an inaccurate prognosis of the disruption length. The prognosis may be updated several times, and consequently, this creates uncertainty for resuming the train service. This shape of resilience curve appears most typical for section/signal failures, which is intuitive considering that the exact nature of this type of failure could be more difficult to identify than for the other disruption causes.

4.4. Mean and median resilience curve per disruption cause

To obtain a more general view of the resilience curve, the representative (i.e. mean and median) resilience curves are drawn for the studied disruption causes, which are train defects, section/signal failures, collisions, switch failures and overhead line failures. For each disruption cause, the mean and median performance are calculated at each time instant t, where time is expressed as a percentage of the disruption length rather than in minutes. For each disruption, 101 measurements are taken, from $t = 0\%$ to $t = 100\%$. Thus, all resilience curves are normalized along the time axis so they can be presented on the same scale. For each disruption cause, the mean and median curve and the central 80 % range are shown in Fig. 6 (a)-(e) for $\lambda = 0.67$, where the central 80 % range is defined as the range of observations between the 10th and the 90th percentile. The mean curves for all disruption causes are shown together in one plot in Fig. 7 (a)-(c) for composite performance ($\lambda = 0.67$), punctuality ($\lambda = 0$) and traffic intensity ($\lambda = 1$), respectively.

The mean and median resilience curves in Fig. 6 (a)-(e) show that the curves do not necessarily resemble the pronounced shape of a bathtub. In particular, we recognize a nearly symmetrical hammock shape for train defects, and a bathtub shape for section/signal failures, switch failures and overhead line failures. For collisions, the mean and median curve bear a stronger resemblance to the gradual recovery curve. Still, the width of the 80 % range across disruption causes shows that an arbitrary resilience curve could deviate significantly from the mean curve. Also, train defects and overhead line failures tend to have a more narrow 80 % range compared to the others disruption causes. Note that the median curves are mostly located higher on the vertical axis than the corresponding mean curves, which suggests that high-impact disruptions may be less common than low-impact disruptions. The width of the 80 % range is relatively small for train defects, which suggests that train defects are the most consistent type of disruption. Instead, the width of the 80 % range for the infrastructure related causes is largest in the transition from first to second phase or early in the second phase, which suggests that those disruptions are more heterogeneous in terms of degradation and response behavior than in terms of recovery behavior. Finally, the width of the 80 % range for collisions is fairly constant from $t = 20\%$ until $t = 60\%$. It shows that certainly not all collisions have a gradual recovery curve, and that a significant number of collisions may have a resilience curve similar to the bathtub shaped curve or the steady state curve with a distinguishable second phase.

The resilience curves in Fig. 6 also suggest that differences in the performance dynamics exist among disruptions of varying causes. In particular, train defects seem to be the least impactful disruptions on average (Fig. 6 (a)), whereas collisions seem to be the most impactful disruptions on average (Fig. 6 (c)). This is most obvious in terms of traffic intensity (Fig. 7 (c)), as it appears that train defects cause relatively few cancellations on average, while collisions cause relatively many cancellations on average. Switch failures also cause relatively few cancellations on average compared to the other infrastructure related causes (i.e. section/signal failures and overhead line failures). Although the differences in terms of punctuality are smaller (Fig. 7 (b)), it is observed that punctuality is affected more strongly on average for the infrastructure related causes than for train defects and collisions, particularly at the start of the disruption. This may be explained by the fact that train drivers can be instructed to drive past the failure location at reduced speed to see if the infrastructure failure disappears by itself, which does not apply to train defects and collisions. The resulting delays can escalate quickly, especially on busy routes, which causes the lower punctuality.

4.5. Comparison of resilience metrics across disruption causes

Group comparisons of the resilience metrics are performed to determine if differences exist among disruptions of varying causes. Table 3 gives descriptive statistics of the resilience metrics: (1) averaged over all disruptions, and (2) per disruption cause. It includes the number of disruptions and the mean, median and standard deviation (SD) of the resilience metrics. The table shows that a single disruption on average has a degradation time of 49.55 min, a response time of 79.17 min, a recovery time of 70.61 min, a maximum

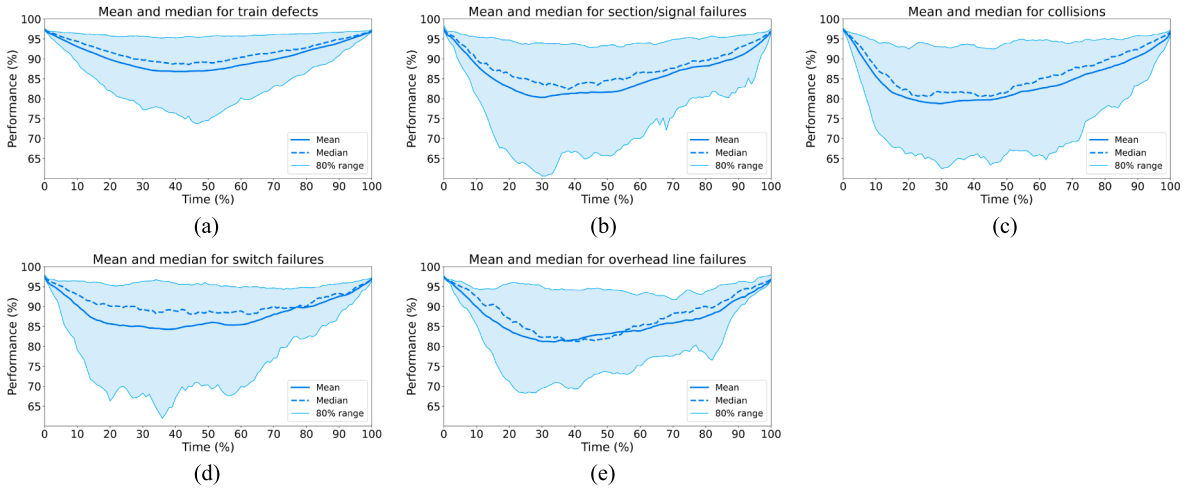


Fig. 6. Mean and median resilience curve per disruption cause.

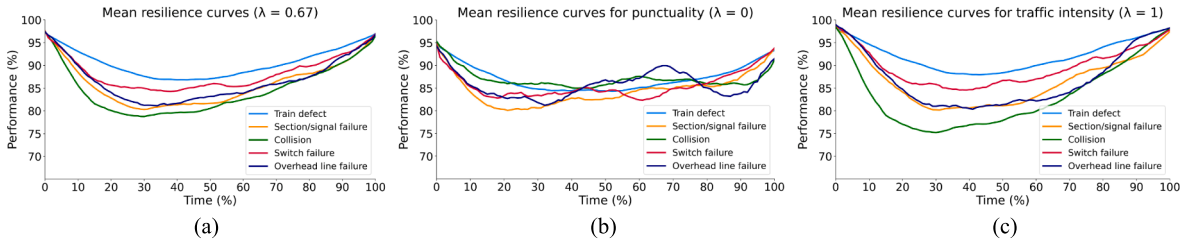


Fig. 7. Mean resilience curve for different values of lambda.

impact of 18.09 %, a performance loss of 2,096 min, a degradation profile of −19.54 % and a recovery profile of.

−54.46 %. However, the average over all disruptions is not necessarily representative for each disruption cause. For example, there is a large difference in the average response time between train defects (53.91 min) and overhead line failures (112.13 min). Also, note that the standard deviations of the metrics are relatively large compared to the means. This suggests that the performance dynamics are quite heterogeneous. For example, the mean degradation time for switch failures equals 69.88 min, while the standard deviation equals 84.32 min, indicating that there may have been a number of switch failures with a long, low-impact first phase. Given the size of the standard deviations, the mean and median degradation profile and recovery profile are relatively close to zero for all disruption causes, which indicates that the shape of the resilience curve in the transition phases is neither strongly concave nor strongly convex on average, but rather linear or mixed.

Since the disruption cause is taken as the variable that defines group membership, there are five groups and seven comparisons to be made: one for each resilience metric. Group comparisons are by definition two-sided tests. Therefore, the null hypothesis H_0 and alternative hypothesis H_1 are as follows:

H_0- The mean of the resilience metric is the same for each disruption cause, and;
H_1- The mean of the resilience metric is different per disruption cause.

The test results of Welch's ANOVA are presented in Table 4, which provides the F-statistic, p-value and η-squared, and states whether or not the null hypothesis is rejected at $\alpha = 0.05$. In saying that the null hypothesis is rejected, it is assumed that all other assumptions in the statistical model are correct, since a small p-value "simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct" (Greenland et al., 2016). The effect size η-squared explains which part of the variation in the dependent variable is associated with group membership (Lakens, 2013). As a rule of thumb, η-squared = 0.01 is considered small, η-squared = 0.06 is considered medium and η-squared = 0.14 is considered large. The results in Table 4 show that the first five metrics are significantly different per disruption cause, and that the effects are medium to large. The largest effect size, η-squared = 0.169, is obtained with regard to the performance loss. In contrast, the degradation profile and the recovery profile show insignificant differences, with a low η-squared equal to 0.012 and 0.008, respectively, and thus, the null hypothesis is not rejected for these two resilience metrics.

In addition to Welch's ANOVA, the Games-Howell post hoc test is performed to assess the differences in the resilience metrics between any two groups of disruptions. The most telling results of the Games-Howell test are presented in Table 5, which provides the groups A and B, the difference in their means, the standard error (SE), t-value, p-value, Hedges' g and common language effect size (CLES). The effect size Hedges' g expresses the difference between the means of two groups as a proportion of the standard deviation of

Table 3
Descriptive statistics of the resilience metrics per disruption cause for $\lambda = 0.67$.

		DT (minutes)			RST (minutes)			RCT (minutes)			MI (%)		
Disruption cause	N	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Train defect	202	40.62	35.00	25.03	53.91	28.00	52.23	56.79	47.00	41.75	13.24	10.81	10.47
Section/signal failure	97	53.61	43.00	44.53	91.04	67.00	79.75	78.60	72.00	57.37	22.35	17.80	13.50
Collision	96	55.96	40.00	50.08	110.01	99.00	69.22	93.33	82.00	63.86	23.90	20.55	13.91
Switch failure	34	69.88	47.50	84.32	92.79	69.50	73.81	76.24	53.50	58.73	17.89	13.23	13.44
Overhead line failure	16	56.13	49.50	28.98	112.13	110.00	27.77	48.25	47.00	25.30	19.03	16.78	11.86
Average	445	49.55	40.00	43.45	79.17	57.00	68.16	70.61	58.00	53.70	18.09	14.68	13.06

		PL (minutes)			DP (%)			RP (%)		
Disruption cause	N	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Train defect	202	1190.34	736.17	1590.88	−10.86	0.82	86.67	−52.90	−4.77	149.68
Section/signal failure	97	2653.90	2153.74	2214.11	−24.66	−9.39	282.33	−43.07	−0.77	223.70
Collision	96	3375.52	2889.97	2374.67	−10.60	−0.37	220.27	−41.10	−22.48	359.17
Switch failure	34	2159.73	1308.50	2185.02	−109.25	−0.94	656.15	−124.36	0.59	312.87
Overhead line failure	16	2358.22	1772.25	2091.90	38.82	−20.42	221.87	−74.74	−36.20	127.39
Average	445	2096.83	1423.82	2170.91	−19.54	0.00	255.88	−54.46	−5.37	238.42

Table 4
Welch's ANOVA test results for $\lambda = 0.67$.

Resilience metric	F-statistic	p-value	η-squared	H_0 rejected ($\alpha = 0.05$)
Degradation time	4.770	1.77E-03	0.043	Yes
Response time	22.352	1.06E-12	0.125	Yes
Recovery time	9.824	1.45E-06	0.081	Yes
Maximum impact	15.954	1.53E-09	0.129	Yes
Performance loss	21.533	6.99E-12	0.169	Yes
Degradation profile	0.437	7.82E-01	0.012	No
Recovery profile	0.611	6.56E-01	0.008	No

this difference. As a rule of thumb, g = 0.2 is considered small, g = 0.5 is considered medium and g = 0.8 is considered large (Cohen, 1988). Although it is generally preferred not to use the rules of thumb and instead compare the effect sizes to earlier results in similar research (Lakens, 2013), such results are not available in this case. The other effect size, CLES, expresses the probability that a randomly sampled observation from one group will have a higher measurement value than a randomly sampled observation from another group.

The results in Table 5 show that the largest differences in the resilience metrics mostly relate to comparisons where train defects are less impactful and/or collisions are more impactful than the other group. For example, train defects have a significantly shorter response time than all other causes, and also, they have the smallest maximum impact and performance loss. Collisions on the other hand have the highest maximum impact and performance loss. Collisions also have the longest recovery time, which is significantly longer than for train defects and overhead line failures. Medium-sized differences involving overhead line failures (i.e. for the degradation time, response time, recovery time and performance loss) are not found to be significant at $\alpha = 0.05$, which may be explained by the small group size. To conclude, the test results are consistent with the mean resilience curves presented in Fig. 6, which already suggested that train defects are the least impactful and collisions are the most impactful disruptions.

The values of the resilience metrics and the ANOVA results are verified for other values of λ which put at least half the weight on traffic intensity and are relatively easy to communicate, namely $\lambda = 0.50$, $\lambda = 0.75$, $\lambda = 0.80$ and $\lambda = 1$. As λ increases, it is observed that performance loss steadily increases for collisions, while it steadily decreases for switch failures. The maximum impact of collisions also increases with increasing λ, which underlines the disruptive effect of collisions in terms of traffic intensity. For $\lambda = 1$, the degradation time and the recovery time are shorter on average than for smaller values of λ, which means the transition phases (and the disruption as a whole) are observed to last shorter when delays are not accounted for.

The results of Welch's ANOVA for the selected values of λ are summarized in Table 6, which presents the effect size η-squared per resilience metric. As λ increases, the differences between groups become more obvious for the response time, maximum impact and performance loss. On the contrary, the differences between groups for $\lambda = 1$ are less obvious for the degradation time and the recovery time. Note that there are some inconsistencies in the increasing or decreasing trend, for example with $\lambda = 0.80$. One possible explanation for these inconsistencies is that there may be a number of resilience curves which are sensitive to changes in λ due to strong fluctuations in punctuality and/or traffic intensity throughout the disruption. Consequently, the timepoints could change significantly for small changes in λ. Thus, $\lambda = 0.80$ might be a particularly unlucky parameter value for a small subset of disruptions. A second explanation is that there are slight differences in the disruption samples, since the sample depends on the conditions outlined in Section 4.1. Even though target performance was adjusted for λ, some disruptions do not appear in all samples.

Table 5
Games-Howell test results for $\lambda = 0.67$ and |Hedges' g| ≥ 0.5.

Metric	Group A	Group B	(A - B)	SE	t-value	p-value	Hedges' g	CLES
DT	Overhead line failure	Train defect	15.51	7.46	2.079	0.274	0.538	0.649
RST	Collision	Train defect	56.10	7.96	7.046	0.001	0.871	0.732
RST	Overhead line failure	Section/signal failure	21.08	10.67	1.977	0.289	0.530	0.647
RST	Overhead line failure	Train defect	58.22	7.86	7.410	0.001	1.918	0.913
RST	Section/signal failure	Train defect	37.14	8.89	4.176	0.001	0.515	0.642
RST	Switch failure	Train defect	38.89	13.18	2.950	0.040	0.545	0.651
RCT	Collision	Overhead line failure	45.08	9.08	4.964	0.001	1.331	0.828
RCT	Collision	Train defect	36.54	7.15	5.112	0.001	0.632	0.673
RCT	Overhead line failure	Section/signal failure	−30.35	8.60	−3.529	0.008	−0.946	0.250
RCT	Overhead line failure	Switch failure	−27.99	11.89	−2.353	0.146	−0.702	0.307
MI	Collision	Train defect	10.66	1.60	6.663	0.001	0.824	0.720
MI	Section/signal failure	Train defect	9.11	1.56	5.854	0.001	0.721	0.695
PL	Collision	Switch failure	1215.80	446.27	2.724	0.062	0.541	0.650
PL	Collision	Train defect	2185.18	266.96	8.185	0.001	1.012	0.763
PL	Overhead line failure	Train defect	1167.88	534.82	2.184	0.234	0.565	0.656
PL	Section/signal failure	Train defect	1463.56	251.13	5.828	0.001	0.718	0.695

Table 6
η-squared per resilience metric for different values of the performance weight λ.

λ	η^2(DT)	η^2(RST)	η^2(RCT)	η^2(MI)	η^2(PL)	η^2(DP)	η^2(RP)
0.50	0.047	0.129	0.068	0.099	0.137	0.011	0.008
0.67	0.043	0.125	0.081	0.129	0.169	0.012	0.008
0.75	0.039	0.146	0.069	0.144	0.186	0.017	0.005
0.80	0.051	0.154	0.063	0.152	0.196	0.021	0.002
1.00	0.028	0.188	0.032	0.151	0.194	0.018	0.014

5. Discussion

The results of the case study show that the performance dynamics are quite heterogeneous. Even though differences exist among disruptions of varying causes in terms of the resilience metrics, and specific shapes of resilience curves are found to be more typical for certain disruption causes, there is still significant heterogeneity among disruptions within each group. Since the composite performance indicator represents the interaction between punctuality and traffic intensity, this heterogeneity also demonstrates the complexity of the interaction between the two individual performance indicators. It is frequently observed that punctuality drops at the start of the disruption, before any trains are canceled. Later, when trains are canceled or short-turned towards the second phase, punctuality might recover again, because a lower number of trains operating means that delays can propagate less easily. If the train service is then resumed too fast or in a way that is infeasible for the TOCs, punctuality might drop again while traffic intensity recovers. Alternatively, it could be the case that punctuality is unaffected because trains are canceled immediately, or that punctuality and traffic intensity recover at a similar rate. Plotting the resilience curve for the individual performance indicators as well as for different combinations of them (i.e. for different values of λ) created a foundation to better understand this interaction.

The fact that the resilience curve can have different shapes, even though each curve results from the same largely predefined and standardized process, means that it would be worth discussing the potential of establishing a preferred resilience curve behavior. Related to that, one could assess the consequences of a resilience curve that is different from the theoretical bathtub shape. In essence, the bathtub model tends to be mainly useful from a conceptual point of view, as it helps translating the message that performance is temporarily lower than usual. In practice, however, the real-time conditions in the network determine the actual operational capabilities, and these conditions could change throughout the disruption. If the resilience curve were to follow the shape of a bathtub, which it occasionally does, this could create predictability towards the TOCs and the passengers. In particular, maintaining a steady level of performance during the second resilience phase could lead to more accurate constraints for the TOCs for rescheduling their rolling stock and crew (e.g. required short-turning locations, number of trains remaining in operation). This might make it easier for TOCs to adjust their operations to the contingency plan and guarantee the availability of rolling stock and crew at the prognosed start of recovery. As a result, passengers would know they could rely on the revisited timetable to reach their destination in spite of the disruption.

Regardless of the type of disruption, an array of factors might affect how system performance develops during a disruption, and as a result, affect the shape of the resilience curve and the values of the resilience metrics. Based on expert judgment and interviews with practitioners, such factors may be categorized as characteristics of the infrastructure, timetable, human factor, information supply or external conditions. An overview of explanatory factors per category is presented in Table 7. The factors in this table could explain part of the variation in the range of resilience curves observed in Fig. 6 (a)-(e). Determining the magnitude of the effect of each factor would require further investigating. The number of explanatory factors alone demonstrates that each disruption could be treated as a unique case. Still, the results of the resilience quantification show that general conclusions can be drawn by studying all of these unique disruptions simultaneously, and that data-driven approaches have the potential to evaluate and improve resilience, even when part of the data are confounded by human interference.

6. Conclusion

In this paper, a data-driven resilience quantification approach was proposed based on a newly developed framework. The approach involves collecting traffic realization data for a large and heterogeneous set of disruptions and reconstructing the resilience curves for these disruptions. Each resilience curve was described by a set of resilience metrics that were evaluated in group comparisons. The five most common disruption causes were studied, which are train defects, section/signal failures, collisions, switch failures and overhead line failures. By specifying the disruption cause as the grouping variable, differences could be identified among disruptions of varying causes in terms of the resilience metrics.

The approach was applied to a case study of the Dutch railway network. The main conclusion based on the results is that the system performance of a railway network during a disruption may approximately follow the shape of the resilience curve as depicted in theory. However, there is significant heterogeneity in the resilience curve behavior, even within disruptions of the same cause. In total, eight distinctive shapes of resilience curves could be recognized. Some resilience curves are fairly well behaved: they degrade, remain steady for some time and recover again, while other resilience curves may show atypical behavior and could be quite unpredictable. Also, different representative curves were obtained per disruption cause, which tends to uncover a structurally different system behavior dependent on the disruption cause. With regard to the resilience metrics, significant differences were observed among disruptions of varying causes in terms of the degradation time, response time, recovery time, maximum impact and performance loss.

Table 7
Factors per category that might affect how performance develops during a disruption.

Category	Explanatory factors
Infrastructure	Number of railway tracks, number of railway switches, network connectivity
Timetable	Number of train series, train frequency, ratio of intercity traffic versus regional traffic, number and length of freight trains
Human factor	Experience of the involved actors, proactive attitude of traffic controllers, change in workload, time pressure to reach the second phase, time pressure to resume the train service
Information supply	Swiftness of reporting the disruption, swiftness of communication throughout the chain, clarity about the cause and location, completeness of information availability of a contingency plan, certainty about the prognosed end time, knowledge within the infrastructure manager of rolling stock and crew rescheduling by the TOC
External conditions	Time of day, weather conditions

The largest effect size was obtained for the performance loss. Post hoc tests showed that train defects are the least impactful and collisions are the most impactful disruptions on multiple resilience metrics. This finding is consistent with the mean resilience curves that were reconstructed for the various disruption causes.

The resilience quantification discussed in this paper has some limitations. Most importantly, less than one third of the initial 1,541 disruptions matching the top five causes were eventually studied as single disruptions. The studied disruptions are mainly the shorter ones, since those have a lower chance of being affected by other, simultaneous disruptions. Furthermore, it was assumed that all delays and cancellations observed in the studied impact area were related to the disruption, while this does not necessarily have to be the case. Also, scheduled maintenance works were not accounted for, while these could have reduced the infrastructure capacity in some disruptions more than usual, potentially adding to the disruptive effects.

Several promising future research directions could be considered. First, the interaction between punctuality and traffic intensity could be studied in more detail. In particular, their interplay within the composite indicator could be explored to understand to what extent and under which conditions changes in one performance indicator could affect the other indicator. Second, it would be interesting to assess resilience during the COVID-19 pandemic and in the post-pandemic period, to understand potential changes in the performance dynamics due to reduced levels of passenger demand and train services. Third, the data-driven resilience quantification could be performed for railway networks in other countries to determine the practical differences between anticipation-based (e.g. the Netherlands, Austria) and reaction-based (e.g. Belgium, Denmark) disruption management, and to determine the pros and cons of these approaches based on realization data. In some countries, more visible fluctuations in performance during the day may be observed, and thus, their impact on resilience would be worth exploring in the future. And more, the quantification approach could be extended to other types of networks, such as other public transport modes and air transport. Fourth, the quantification approach could be extended to investigate multiple, simultaneous disruptions, with possible modifications to the resilience metrics and to the identification of the start and end time of a disruption. Fifth, a more detailed analysis could be performed on the relationships between the shapes of resilience curves and the disruption causes, to determine the degree in which each shape is represented in total and per disruption cause. Similarly, more robust DP and RP metrics are needed to capture more complex behaviour of the system, i.e. non-concave/convex. Finally, as seen in Bešinović et al. (2022), certain disruptions can have more implications on passengers than on services, and vice versa. In the future, with more realized passenger movement data available, we shall aim towards passenger-focused resilience analysis. This would lead to further exploring the realized relations between impacts on passengers versus services in real-life railway operations.

Overall, the developed data-driven resilience quantification approach, with further improvements, can represent a useful decision support tool for practitioners to provide an in-depth understanding of the performance dynamics, determine gaps in the current state of the practice, and recognize specific types of disruptions and/or specific phases in a disruption which deserve attention to improve disruption management.

CRediT authorship contribution statement

Max J. Knoester: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization, Validation, Writing – original draft. **Nikola Bešinović:** Conceptualization, Methodology, Writing – review & editing. **Amir Pooyan Afghari:** Conceptualization, Methodology, Writing – review & editing. **Rob M.P. Goverde:** Conceptualization, Methodology, Writing – review & editing. **Jochen van Egmond:** Conceptualization, Resources.

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.

Acknowledgements

The authors thank the colleagues in the Traffic Management Division of ProRail for their support, openness and detailed discussions, with a special thanks to Wilco Tielman for supplying the necessary realization data. This research is partly supported by the

SORTEDMOBILITY project, which is supported by the European Commission and funded under the Horizon 2020 ERA-NET Cofund scheme under grant agreement N° 875022.

Appendix A. Breadth first search procedure (single start vertex)

A custom breadth first search algorithm was developed in which the start vertex or vertices is/are the boundary point(s) of a disruption, and the visited vertices are the timetable points in the first and second impact area. The custom breadth first search algorithm for single start vertex is as follows:

1. Add the start vertex to an empty queue;
2. While the queue contains elements, take the first node from the queue;
3. Search for neighbors of this node while keeping track of the parent and the visited nodes, including their impact area;
4. Raise the impact area of the neighbor by 1 if its parent node is a decoupling point;
5. Add the neighbor to the queue, except for when it is a second decoupling point or it is already in the queue;
6. Repeat steps 2–5 until the queue is empty.

For two and three start vertices, the BFS follows the same structure and is further extended, details can be found in Knoester (2021).

Appendix B. Steady state detection procedure

The modified steady state detection algorithm was developed to detect the steady state in the resilience curve, and thus, the start and end point of the second resilience phase - i.e. the response phase. The pseudocode of the modified steady state detection algorithm is as follows:

1. Specify the datapoints;
2. Define tuning parameters;
3. Search for a starting point in the lower performance range;
4. Specify the subset of data for the corresponding time window;
5. Perform linear regression analysis on this subset of data;
6. Mark the window as steady if the regression model is statistically significant;
7. Increment the start and end of the time window by 1;
8. Repeat steps 4–7 until the regression model is no longer significant and performance is above the search threshold.

References

ACM. (2019, March 22). ACM Rail Monitor: the Netherlands has Europe's busiest railway network. https://www.acm.nl/en/publications/acm-rail-monitor-netherlands-has-europes-busiest-railway-network.
Adjetey-Bahun, K., Birregah, B., Châtelet, E., Planchet, J.L., 2016. A model to quantify the resilience of mass railway transportation systems. Reliab. Eng. Syst. Saf. 153, 1–14.
Bababeik, M., Khademi, N., Chen, A., 2018. Increasing the resilience level of a vulnerable rail network: The strategy of location and allocation of emergency relief trains. Transp. Res. Part E: Logist. Transp. Rev. 119, 110–128.
Bešinović, N., 2020. Resilience in railway transport systems: a literature review and research agenda. Transp. Rev. 40 (4), 457–478.
Bešinović, N., Wang, Y., Zhu, S., Quaglietta, E., Tang, T., Goverde, R.M., 2021. A Matheuristic for the Integrated Disruption Management of Traffic, Passengers and Stations in Urban Railway Lines. IEEE Trans. Intell. Transp. Syst.
Bešinović, N., Nassar, R.F., Szymula, C., 2022. Resilience assessment of railway networks: Combining infrastructure restoration and transport management. Reliab. Eng. Syst. Saf. 224, 108538.
Bruneau, M., Chang, S.E., Eguchi, R.T., Lee, G.C., O`Rourke, T.D., Reinhorn, A.M., Shinozuka, M., Tierney, K., Wallace, W.A., Von Winterfeldt, D., 2003. A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities. Earthq. Spectra 19 (4), 733–752.
Büchel, B., Spanninger, T., Corman, F., 2020. Empirical dynamics of railway delay propagation identified during the large-scale Rastatt disruption. Sci. Rep. 10, 18584.
Cacchiani, C., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., Wagenaar, J., 2014. An overview of recovery models and algorithms for real-time railway rescheduling. Transp. Res. B Methodol. 63, 15–37.
Cats, O., Jenelius, E., 2014. Dynamic Vulnerability Analysis of Public Transport Networks: Mitigation Effects of Real-Time Information. Netw. Spat. Econ. 14, 435–463.
Chan, R., Schofer, J.L., 2016. Measuring Transportation System Resilience: Response of Rail Transit to Weather Disruptions. Nat. Hazard. Rev. 17 (1), 05015004.
Chen, Z., Wang, Y., 2019. Impacts of severe weather events on high-speed rail and aviation delays. Transp. Res. Part D: Transp. Environ. 69, 168–183.
Cimellaro, G.P., Reinhorn, A.M., Bruneau, M., 2010. Seismic resilience of a hospital system. Struct. Infrastruct. Eng. 6 (1–2), 127–144.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Routledge.
Dalheim, Ø.Ø., Steen, S., 2020. A computationally efficient method for identification of steady state in time series data from ship monitoring. J. Ocean. Eng. Sci. 5 (4), 333–345.
Dekker, M.M., Panja, D., 2021. Cascading dominates large-scale disruptions in transport over complex networks. PLoS One 16 (1), e0246077.
D'Lima, M., Medda, F., 2015. A new measure of resilience: An application to the London Underground. Transp. Res. A Policy Pract. 81, 35–46.
Dorbritz, R., 2011. Assessing the resilience of transportation systems in case of large-scale disastrous events. 8th International Conference on Environmental Engineering (ICEE) Selected Papers, 1070-1076.

Ghaemi, N., Cats, O., Goverde, R.M.P., 2017. Railway disruption management challenges and possible solution directions. Publ. Transp. 9, 343–364.
Gonçalves, L.A.P.J., Ribeiro, P.J.G., 2020. Resilience of urban transportation systems. Concept, characteristics, and methods. J. Transp. Geogr. 85, 102727.
Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. 31, 337–350.
Hosseini, S., Barker, K., Ramirez-Marquez, J.E., 2016. A review of definitions and measures of system resilience. Reliab. Eng. Syst. Saf. 145, 47–61.
Janić, M., 2015. Reprint of "Modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event". Transp. Res. A Policy Pract. 81, 77–92.
Janić, M., 2018. Modelling the resilience of rail passenger transport networks affected by large-scale disruptive events: the case of HSR (high speed rail). Transportation 45 (2), 1101–1137.
Janić, M., 2019. Modeling the resilience of an airline cargo transport network affected by a large scale disruptive event. Transp. Res. Part D: Transp. Environ. 77, 425–448.
Knoester, M.J., 2021. A data-driven approach for evaluating the resilience of railway networks. Delft University of Technology. Master thesis.
Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863.
Liu, E., Barker, K., Chen, H., 2022. A multi-modal evacuation-based response strategy for mitigating disruption in an intercity railway system. Reliab. Eng. Syst. Saf. 108515.
Liu, K., Zhu, J., Wang, M., 2021. An event-based probabilistic model of disruption risk to urban metro networks. Transp. Res. A Policy Pract. 147, 93–105.
Lu, Q.C., 2018. Modeling network resilience of rail transit under operational incidents. Transp. Res. A Policy Pract. 117, 227–237.
Madni, A.M., Erwin, D., Sievers, M., 2020. Constructing Models for System Resilience: Challenges, Concepts, and Formal Methods. Systems 8 (3).
Malandri, C., Fonzone, A., Cats, O., 2018. Recovery time and propagation effects of passenger transport disruptions. Physica A 505, 7–17.
Mattsson, L.-S., Jenelius, E., 2015. Vulnerability and resilience of transport systems – A discussion of recent research. Transp. Res. A Policy Pract. 81, 16–34.
Mertens, W., Pugliese, A., Recker, J., 2017. Quantitative Data Analysis: A Companion for Accounting and Information Systems Research. Springer.
Moore, E.F., 1959. The shortest path through a maze. Proceedings of the International Symposium on the Theory of Switching, 285-292.
Munoz, A., Dunbar, M., 2015. On the quantification of operational supply chain resilience. Int. J. Prod. Res. 53 (22), 6736–6751.
Nicholson, G.L., Kirkwood, D., Roberts, C., Schmid, F., 2015. Benchmarking and evaluation of railway operations performance. J. Rail Transp. Plann. Manage. 5 (4), 274–293.
Ouyang, M., Dueñas-Osorio, L., Min, X., 2012. A three-stage resilience analysis framework for urban infrastructure systems. Struct. Saf. 36–37, 23–31.
Parkinson, H.J., Bamford, G., 2017. A journey into railway digitisation. Stephenson Conference: Research for Railways 2017, 333–340.
Ren, X., Yin, J., & Tang, T., 2020. Quantitative analysis for resilience-based urban rail systems: A hybrid knowledge-based and data-driven approach. Proceedings of the 29th European Safety and Reliability Conference, 3531–3538.
Schipper, D., Gerrits, L., 2018. Differences and similarities in European railway disruption management practices. J. Rail Transp. Plann. Manage. 8 (1), 42–55.
Spiegler, V.L.M., Naim, M.M., Wikner, J., 2012. A control engineering approach to the assessment of supply chain resilience. Int. J. Prod. Res. 50 (21), 6162–6187.
Uday, P., Marais, K., 2015. Designing Resilient Systems-of-Systems: A Survey of Metrics, Methods, and Challenges. Syst. Eng. 18 (5), 491–510.
Van Aken, S., Bešinović, N., Goverde, R.M.P., 2017. Designing alternative railway timetables under infrastructure maintenance possessions. Transp. Res. B Methodol. 98, 224–238.
Wong, A., Tan, S., Chandramouleeswaran, K.R., Tran, H.T., 2020. Data-driven analysis of resilience in airline networks. Transp. Res. Part E: Logist. Transp. Rev. 143, 102068.
Woodburn, A., 2019. Rail network resilience and operational responsiveness during unplanned disruption: A rail freight case study. J. Transp. Geogr. 77, 59–69.
Xu, Z., Chopra, S.S., 2022. Network-based Assessment of Metro Infrastructure with a Spatial–temporal Resilience Cycle Framework. Reliab. Eng. Syst. Saf. 223, 108434.
Yin, J., Ren, X., Liu, R., Tang, T., Su, S., 2022. Quantitative analysis for resilience-based urban rail systems: A hybrid knowledge-based and data-driven approach. Reliab. Eng. Syst. Saf. 219, 108183.
Zhang, Y., Ng, S.T., 2022. Robustness of urban railway networks against the cascading failures induced by the fluctuation of passenger flow. Reliab. Eng. Syst. Saf. 219, 108227.
Zhou, L., Chen, Z., 2020. Measuring the performance of airport resilience to severe weather events. Transp. Res. Part D: Transp. Environ. 83, 102362.
Zhou, Y., Wang, J., Yang, H., 2019. Resilience of Transportation Systems: Concepts and Comprehensive Review. IEEE Trans. Intell. Transp. Syst. 20 (12), 4262–4276.
Zhu, Y., Ozbay, K., Xie, K., Yang, H., 2016. Using Big Data to Study Resilience of Taxi and Subway Trips for Hurricanes Sandy and Irene. Transp. Res. Rec.: J. Transp. Res. Board 2599 (1), 70–80.
Zhu, Y., Xie, K., Ozbay, K., Zuo, F., Yang, H., 2017. Data-Driven Spatial Modeling for Quantifying Networkwide Resilience in the Aftermath of Hurricanes Irene and Sandy. Transp. Res. Rec.: J. Transp. Res. Board 2604 (1), 9–18.
Zilko, A.A., Kurowicka, D., Goverde, R.M.P., 2016. Modeling railway disruption lengths with Copula Bayesian Networks. Transp. Res. Part C: Emerg. Technol. 68, 350–368.
Zobel, C.W., 2011. Representing perceived tradeoffs in defining disaster resilience. Decis. Support Syst. 50 (2), 394–403.