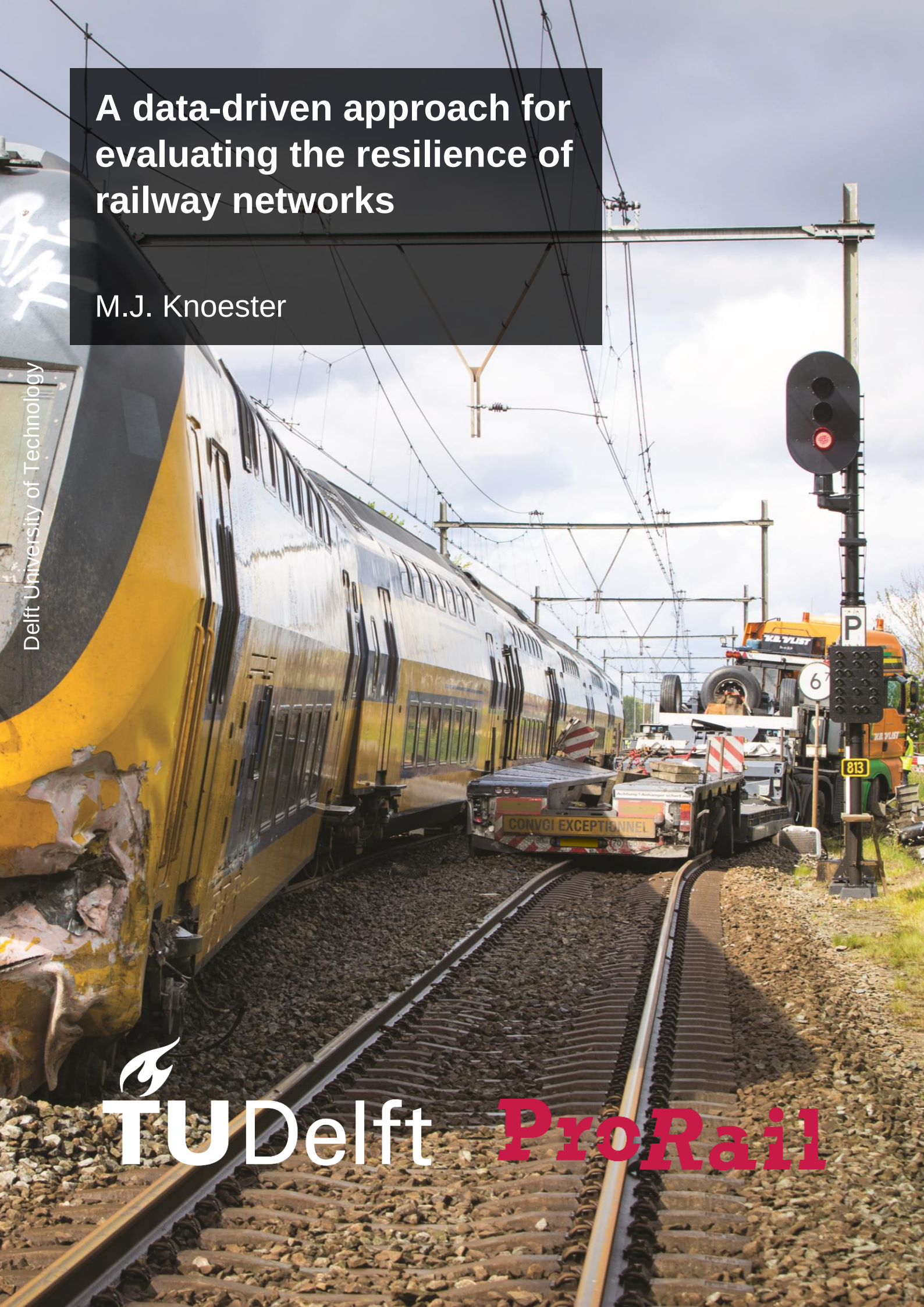
A data-driven approach for evaluating the resilience of railway networks

M.J. Knoester

Delft University of Technology



TU Delft ProRail

Page intentionally left blank

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

by

M.J. (Max) Knoester

to obtain the degree of Master of Science
in Transport, Infrastructure & Logistics
at the Delft University of Technology,
to be defended publicly on Friday October 15, 2021 at 16:00

October 1, 2021

Student no.: 4222539

Committee:	Prof. dr. R.M.P. Goverde	TU Delft, Department of Transport and Planning
	Dr. N. Bešinović	TU Delft, Department of Transport and Planning
	Dr. A.P. Afghari	TU Delft, Safety and Security Science Section
	J. van Egmond	ProRail, Traffic Management Staff

Cover photo by Stefan Verkerk

TU Delft ProRail

Page intentionally left blank

Preface

This thesis report marks the end of a nine year learning journey at TU Delft. It has been a long road for sure, a rocky road at times. But with the years came perseverance, and in this report, I see the crown to my work during the master Transport, Infrastructure and Logistics which I started in September 2018.

Although railways were already the topic of my bachelor thesis, it was not until the TIL design project that I really started to develop an interest in this field. That is also where I met Nikola, one of the daily supervisors from TU Delft, who I would like to thank for his enthusiasm and constructive criticism. Naturally I would also like to thank my other daily supervisor Amir, in particular for helping me create a clear and consistent storyline, and my professor Rob, for overseeing the entire process. All of their feedback has helped me strive for the best possible result within the available time and research scope. What I appreciated most of all was the level of detail in the feedback which, as we have discussed, was a natural result of the effort I had put in myself. The fact that we were able to go into the details also indicates there were never any major roadblocks along the way, even though I deliberately did some things the hard way, for example by developing a dedicated graph search algorithm for three start vertices. Looking back, I believe the results were worth the effort.

When I first contacted Nikola about this thesis opportunity in December 2020, it was not clear yet that I would do the thesis in ProRail. Luckily the draft proposal reached Jochen, who saw this as an interesting research opportunity and thus became my company supervisor. Working as an intern in ProRail has been a pleasant and valuable experience, even though I have spent most of the time in my dorm room due to COVID-19 restrictions. Still, the VGB team members made me feel welcome and part of the team, which is why I regret having to leave them so soon again. My gratitude goes out to all of them. In particular, I would like to thank Jochen for his support, enthusiasm and sincerity, and for trying to help me find a pace that can be sustained in the long term. Also, I would like to thank the colleagues of the PAB for helping me get up to speed with Sherlock and obtain the necessary data, and all of the other colleagues that I spoke to for their engagement. Hearing how they identified with the results was probably one of the biggest compliments I could wish for.

It would feel as a great reward if the findings from this thesis will eventually help improve disruption management practices in the Netherlands. I believe ProRail is already heading in the right direction, however, they just need to take the next step. Even if this report ends up on a shelf or in a drawer, I will at least have drawn attention to the topic of resilience and shown that much more can be done with the currently available data. I enjoyed exploring Sherlock a lot, and although it already contains a wealth of information, I hope the issues that I raised over the course of the research will benefit future users.

Last but not least, I would like to thank my parents for their unconditional support, even during the times in my bachelor when progress was slow. Now, it is time for me to make the leap into practice. Though I will abandon the topic of resilience for now, I will remain active in the rail industry. Hopefully you will find that my interest in this field is reflected in the quality of this report. Enjoy reading!

Max Knoester
Delft, October 2021

Summary

Research problem

The Dutch railway network is one of the busiest in Europe. Under normal conditions, trains run according to the timetable and minor variations in the train service may occur. However, as in any type of transportation system, it is inevitable that disruptions in the network will occur. Disruptions are unexpected changes that are caused by the failure of infrastructure or rolling stock, unscheduled maintenance, extreme weather or other external events. In case of a disruption, adjustments need to be made to the timetable, crew planning and rolling stock planning. The consequence is generally that trains are canceled or experience serious delays. Disruptions can easily propagate through the network, and their effects may even build up to a networkwide scale. In addition, the number of disruptions in the Dutch railway network per year and per train kilometer has been steadily increasing over the last decade.

System performance during disruptions can be visualized in the resilience curve, which is also known as the bathtub model. Three phases are distinguished in the curve: the first and third are transition phases, whereas the second phase represents a disrupted yet stable system. In reality though, the curve has remained mostly theoretical in nature. Not much is known about the shape of the curve or about railway system performance during disruptions in general. This makes it hard to design appropriate measures for disruption management such as contingency plans. More and better quantitative knowledge about resilience would contribute to the effective allocation of resources to improve the design and application of measures, and create a more resilient railway network. Therefore, the following research question was formulated:

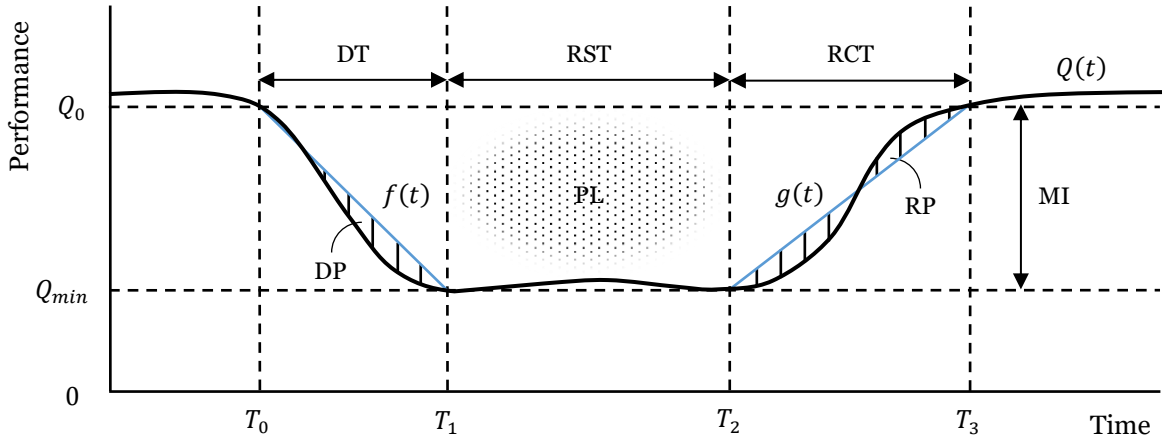
How does the system performance of a railway network develop during disruptions?

Methodology

The research question was answered by applying a combination of methods. First, a literature review was performed which focused on the quantification of railway system performance. The four cornerstones of the review were the resilience definitions, performance indicators, resilience evaluation approaches and resilience metrics. Second, interviews with practitioners were held to learn about the state of the practice regarding disruption management. Third, the main research method was a quantitative data analysis of historical railway traffic realization data. Realization data includes the plan time and realization time of a train activity, which can be an arrival, passing or departure at a certain location. These data were combined with disruption log data from ProRail's data analysis application "Sherlock" to calculate the evolution of system performance during disruptions. Performance was defined herein as a weighted sum of traffic intensity and traffic punctuality, and it was calculated as a centered moving average over a time period of 30 minutes. This resulted in resilience curves such as the schematic example below. The resilience curves were not just determined for the disrupted line or timetable point, but for the entire area where disruption effects are expected to occur. According to the terminology used in ProRail, this area is referred to as the first and second impact area.

Each resilience curve was described quantitatively in terms of seven resilience metrics: the degradation time (DT), response time (RST) and recovery time (RCT), which represent the time dimension; the maximum impact (MI), which represents the performance dimension; and the performance loss (PL), degradation profile (DP) and recovery profile (RP), which represent both dimensions at once. Since performance was calculated over the entire impact area, the spatial aspect of the disruption dynamics was incorporated in the performance dimension. Resilience curves were determined for the top five most common disruption causes where a contingency plan was applied, which are train defects, section/signal failures, collisions, switch failures and overhead line failures. Differences and similarities in the resilience metrics across

disruptions of different causes were evaluated and relationships between resilience metrics were identified in incrementally designed experiments as part of a case study. The case study focused on passenger traffic in the Dutch railway network in timetable year 2019, which was the last regular timetable year before the COVID-19 pandemic. Among others, the experiments involved evaluating an example case, identifying different types of resilience curves, drawing the mean and median resilience curve per disruption cause, evaluating differences between the observed and reported timepoints that mark the end of a phase, performing group comparisons of the resilience metrics and performing regression analysis.



Results

The results of the case study showed that the resilience curve does not necessarily resemble the theoretical shape of the bathtub. Instead, eight different types of resilience curves were identified. Despite the fact that some types of curves appear more characteristic of one disruption cause than of another, there is still significant heterogeneity in the shape of the curve, even among disruptions of the same cause. Regarding the time differences, it was found that disruptions are reported late on average, and also, that disruptions last longer on average than reported. In fact, the third phase is nearly as long as the second phase on average, which makes it much longer than experienced in practice. Based on the mean and median resilience curves, it was postulated that train defects are the least impactful single disruptions overall, whereas collisions are the most impactful single disruptions overall. This assumption seemed intuitive and was confirmed in the group comparisons, which showed that significant differences exist among disruptions of different causes in terms of the first five resilience metrics. The largest differences in general were observed in the response time, maximum impact and performance loss. Compared to the other types of disruptions, collisions stand out negatively in terms of recovery time, maximum impact and performance loss.

Furthermore, a significant positive relationship was identified between performance loss on the one hand and the maximum impact and total duration on the other hand, where the relationship with the maximum impact was the most obvious of the two. This means that the cumulative loss of performance depends more strongly on how much performance is reduced at the lowest point in the curve than on how long the disruption lasts. No relationships were found for the other resilience metrics and additional explanatory variables. The absence of other relationships means that reaching the next phase quickly does not necessarily tell much. More specifically, it indicates that rushing through the first phase does not necessarily result in better (or worse) performance during the remainder of the disruption, in contrast to what seems to be the general belief in ProRail. Instead, it would be better to take the time to devise a structurally feasible traffic plan that limits the chance of secondary delays. By complying with the plan, the predictability towards passengers and train operating companies will increase.

Conclusions

The general conclusion is that the system performance of a railway network during disruptions may approximately follow the shape of the resilience curve as depicted in theory, but this does not always have to be the case. Some resilience curves are fairly well behaved: they degrade, remain steady for some time and recover again, while others may show atypical behavior and can even be quite unpredictable. The shape of the curve might be affected by an abundance of factors, which could be categorized as characteristics of the infrastructure, timetable, human action, information supply or external conditions. With regard to disruption management in the Netherlands, improvements could be achieved in each of the resilience phases, and also in the handling of collisions. Although ProRail is already heading in the right direction, the recommendations made in this report could help realize the transition towards even more predefined and proactive disruption management, which is needed to manage an increasingly busy railway network.

Samenvatting

Onderzoeksprobleem

Het Nederlandse spoornetwerk is één van de drukste netwerken in Europa. Onder normale omstandigheden rijden treinen volgens de dienstregeling en doen zich slechts kleine variaties in de treindienst voor. Zoals in elk vervoersysteem is het echter onvermijdelijk dat er verstoringen optreden. Verstoringen zijn onverwachte veranderingen als gevolg van een defect aan de infrastructuur of het materieel, ongeplande werkzaamheden, extreme weersomstandigheden of andere externe oorzaken. In het geval van een verstoring, hierna “calamiteit” genoemd volgens de formele definitie binnen ProRail, zijn aanpassingen aan de dienstregeling, personeelsplanning en materieelplanning nodig. Het gevolg is doorgaans dat treinen worden opgeheven of vertraging oplopen. De effecten van een calamiteit kunnen zich makkelijk door het netwerk verspreiden, en kunnen in het ergste geval zelfs door het hele netwerk merkbaar zijn. Daarnaast is het aantal calamiteiten in het Nederlandse spoornetwerk per jaar en per gereden treinkilometer het afgelopen decennium gestaag toegenomen.

De systeemprestatie tijdens calamiteiten kan worden gevisualiseerd in de veerkrachtcurve, die ook wel bekend staat als het badkuipmodel. De curve bestaat uit drie fases: de eerste en derde fase zijn overgangsfases, terwijl de tweede fase een verstoord doch stabiel systeem beschrijft. In werkelijkheid heeft de curve vooral een theoretisch karakter. Er is weinig bekend over de exacte vorm van de curve of over de systeemprestatie van het spoornetwerk tijdens calamiteiten in het algemeen. Dit maakt het lastig om passende maatregelen zoals versperringsmaatregelen te ontwerpen. Meer en betere kwantitatieve kennis op het gebied van veerkracht zou bijdragen aan de effectieve toekenning van middelen om het ontwerp en de toepassing van maatregelen te verbeteren, en zo een veerkrachtiger spoornetwerk te creëren. Daarom is de volgende onderzoeksvraag geformuleerd:

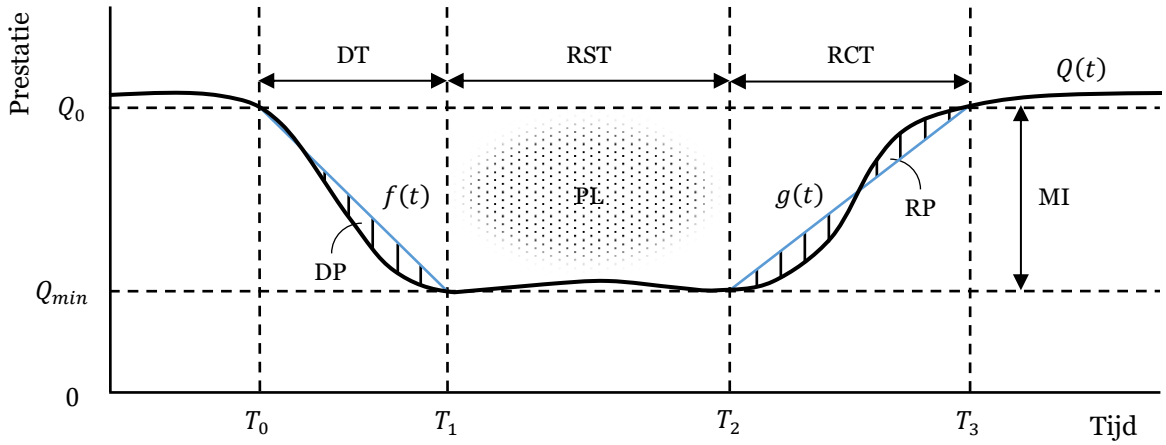
Hoe verloopt de systeemprestatie van een spoornetwerk tijdens calamiteiten?

Methodiek

De onderzoeksvraag is beantwoord door toepassing van een combinatie van onderzoeksmethodes. Ten eerste is een literatuuronderzoek uitgevoerd waarbij de focus lag op het kwantificeren van de systeemprestatie. De vier hoekstenen van het onderzoek waren de veerkrachtdefinities, prestatie-indicatoren, evaluatiemethodes en kengetallen. Ten tweede zijn interviews gevoerd met mensen uit de praktijk om meer te weten te komen over de stand van de praktijk met betrekking tot calamiteitsmanagement. Ten derde is als voornaamste onderzoeksmethode een kwantitatieve data-analyse uitgevoerd van historische realisatiedata. Realisatiedata betreft onder meer de plantijd en realisatietijd van een treinactiviteit. Een activiteit is gedefinieerd als een aankomst, doorkomst of vertrek op een dienstregelpunt. Deze data zijn gecombineerd met de gelogde data uit ProRails data-analyseapplicatie “Sherlock” om het verloop van de systeemprestatie tijdens calamiteiten te berekenen. Prestatie werd hierin gedefinieerd als een gewogen som van verkeersintensiteit en verkeerspunctualiteit en werd berekend als een gecentreerd voortschrijdend gemiddelde over een periode van 30 minuten. Dit heeft geleid tot veerkrachtcurves zoals in het onderstaande, schematische voorbeeld. De curves zijn niet bepaald voor alleen het getroffen baanvak of dienstregelpunt, maar voor het gehele gebied waarin de gevolgen van een calamiteit merkbaar kunnen zijn. Volgens de terminologie binnen ProRail wordt hiernaar verwezen als het eerste en tweede impactgebied.

Elke veerkrachtcurve is kwantitatief beschreven aan de hand van zeven veerkrachtkengetallen: de degradatietijd (DT), handelingstijd (RST) en hersteltijd (RCT), die de tijdsdimensie vertegenwoordigen; de maximale impact (MI), die de prestatiedimensie vertegenwoordigt; en het prestatieverlies (PL), degradatieprofiel (DP) en herstelprofiel (RP), die beide dimensies tege-

lijk vertegenwoordigen. Aangezien de prestatie is berekend over het hele impactgebied, zit het ruimtelijke aspect van de storingsdynamiek opgesloten in de prestatiedimensie. De veerkrachtcurves zijn bepaald voor de vijf meest voorkomende oorzaken waarvoor een versperringsmaatregel is toegedeeld, namelijk defect materieel, sectie-/seinstoringen, aanrijdingen, wisselstoringen en defecte bovenleiding. Verschillen en overeenkomsten in de kengetallen tussen calamiteiten met verschillende oorzaken zijn geëvalueerd, en relaties tussen de kengetallen zijn vastgesteld in incrementeel ontworpen experimenten als onderdeel van een casestudie. De casestudie was gericht op reizigersverkeer in het hele Nederlandse spoornetwerk in dienstregeljaar 2019, het laatste reguliere dienstregeljaar voor de coronapandemie. De experimenten bestonden onder meer uit het uitgebreid evalueren van een enkele calamiteit, het identificeren van verschillende soorten curves, het bepalen van de gemiddelde curve en mediaancurve per storingsoorzaak, het evalueren van de tijdsverschillen tussen de waargenomen en gerapporteerde begin- en eindtijden van de badkuipfases, het uitvoeren van groepsvergelijkingen op basis van de kengetallen en het uitvoeren van regressieanalyses.



Resultaten

De resultaten uit de casestudie lieten zien dat de veerkrachtcurve niet noodzakelijk lijkt op de theoretische vorm van de badkuip. Integendeel, er werden acht verschillende soorten curves herkend. Ondanks het feit dat bepaalde soorten meer typerend lijken voor één storingsoorzaak dan voor een andere oorzaak is er een significante heterogeniteit in de vorm van de curve, zelfs onder calamiteiten met dezelfde oorzaak. Met betrekking tot de tijdsverschillen werd gevonden dat calamiteiten gemiddeld beschouwd te laat worden gemeld, en ook dat calamiteiten gemiddeld beschouwd langer duren dan gerapporteerd. De derde fase duurt gemiddeld beschouwd bijna even lang als de tweede fase, en duurt daarmee aanzienlijk langer dan in de praktijk wordt ervaren. Op basis van de gemiddelde curves en mediaancurves werd verondersteld dat defect materieel de minst impactvolle opzichzelfstaande calamiteiten zijn, terwijl aanrijdingen de meest impactvolle opzichzelfstaande calamiteiten zijn. Deze aanname leek intuïtief en werd bevestigd in de groepsvergelijkingen, waaruit bleek dat er significante verschillen bestaan in de eerste vijf kengetallen tussen calamiteiten met verschillende oorzaken. De grootste verschillen werden gemeten in de handelingstijd, maximale impact en prestatieverlies. Vergeleken met de anders storingsoorzaken sprongen aanrijdingen er in negatieve zin uit als het gaan om de hersteltijd, maximale impact en prestatieverlies.

Verder werd in de regressieanalyse een positief verband gevonden tussen het prestatieverlies enerzijds en de maximale impact en totale duur van een calamiteit anderzijds, waarbij het verband met de maximale impact het meest duidelijk was. Dit verband laat zien dat het cumulatieve verlies in prestatie sterker afhankelijk is van de mate waarin de prestatie is afgenomen op het laagste punt in de curve dan van de totale duur van een calamiteit. Er werden verder

geen verbanden gevonden voor de overige kengetallen en andere verklarende variabelen. De afwezigheid van andere verbanden betekent onder meer dat het snel bereiken van de volgende fase niet per se veel invloed heeft op de prestaties. Meer specifiek betekent dit dat het snel doorlopen van de eerste fase niet per definitie leidt tot een betere (of slechtere) prestatie in het verdere verloop van de calamiteit, in tegenstelling tot wat de heersende gedachte lijkt te zijn binnen ProRail. In plaats daarvan zou het dan ook de voorkeur verdienen om de tijd te nemen voor een structureel maakbaar verkeersplan, zowel voor de tweede als derde fase, waarbij de kans op nieuwe vertragingen zo veel mogelijk wordt beperkt. Door dit plan vervolgens ook na te leven, wordt bovendien de voorspelbaarheid richting de reizigers en de vervoerders vergroot.

Conclusies

De algemene conclusie is dat de systeemprestatie van een spoornetwerk tijdens calamiteiten ongeveer de vorm van de veerkrachtcurve kan volgen zoals deze in de theorie wordt afgebeeld, maar dit hoeft in de praktijk niet altijd het geval te zijn. Sommige curves vertonen redelijk goed gedrag: ze nemen af, blijven een tijd stabiel en herstellen weer, terwijl andere curves atypisch gedrag vertonen en zelfs vrij onvoorspelbaar kunnen zijn. De vorm van de curve zou beïnvloed kunnen worden door een verscheidenheid aan factoren, die gecategoriseerd kunnen worden als eigenschappen van de infrastructuur, dienstregeling, menselijk handelen, informatievoorziening of externe omstandigheden. Met betrekking tot de bijsturingspraktijken in Nederland zijn er in elke badkuipfase, en ook in de afhandeling van aanrijdingen in het algemeen, verbeteringen te behalen. ProRail zit al op de goede weg, maar de aanbevelingen in dit rapport zouden kunnen helpen bij de transitie naar een nog meer vooraf gedefinieerd en proactief calamiteitsmanagement, wat nodig is om het alsmaar drukker wordende spoornetwerk op gang te houden. De Nederlandstalige aanbevelingen zijn te vinden in Appendix B.

Table of contents

List of figures

List of tables

List of abbreviations

Abbreviation	Meaning in Dutch	Meaning in English
A&E	Analyse en Evaluatie	Analysis and Evaluation
AL	Algemeen leider	General controller
ANOVA	Variantieanalyse	Analysis of variance
BBT	Be- en Bijsturing van de Toekomst	Railway Control and Rescheduling of the Future
BFS	N/A	Breadth first search
CMBO	Centraal Monitoring en Beslisorgaan	Central Monitoring and Operations Control Center
DVL	Decentrale verkeersleider	Regional traffic controller
ERTMS	N/A	European Rail Traffic Management System
ETCS	N/A	European Train Control System
HK	Hinderklasse	Hindrance class
IC	Intercity	Intercity
ICB	Incidentenbestrijding	Calamity control
IM	Infrabeheerder	Infrastructure manager
KPI	Kritieke prestatie-indicator	Key performance indicator
MKS	Meldkamer spoor	Control room
NS	Nederlandse Spoorwegen	Dutch Railways
OCCR	Operationeel Controle Centrum Rail	Operational Control Center Rail
OvD-S	Officier van dienst spoor	Duty officer rail
PAB	Prestatie Analyse Bureau	Performance Analysis Bureau
PHS	Programma Hoogfrequent Spoorvervoer	High-Frequency Rail Transport Program
RBC	Regionaal Besturingscentrum	Regional Operations Control Center
RoSA	Rotterdam-Schiphol-Arnhem	N/A
SSD	Steady state detectie	Steady state detection
TAD	Treinafhandelingsdocument	Train delay handling document
TIS	Treinincidentscenario	Train incident scenario
TOC	Spoorwegonderneming	Train operating company
TRDL	Treindienstleider	Train dispatcher
VDB	Verdeelbesluit	Capacity reallocation
VGB	Vooraf Gedefinieerde Bijsturing	Predefined Solutions
VL	Verkeersleiding	Traffic control
VLC	Verkeersleider CMBO	National traffic controller
VSM	Versperringsmaatregel	Contingency plan

1. Introduction

This thesis report explores the dynamics in railway system performance during disruptions. In this first chapter, an outline of the research problem is given. Section 1.1 presents the problem and company context and introduces some basic definitions of disruptions and delays. Section 1.2 presents the problem statement, and Section 1.3 presents the research objectives and approach. Section 1.4 presents the main research question and related subquestions. Section 1.5 describes the research methods, and Section 1.6 delineates the scope of the research. Section 1.7 presents the scientific and societal significance of the research. Finally, Section 1.8 presents a reading guide and a flowchart of the report structure.

1.1. Problem context

The Dutch railway network is known as the busiest in Europe (ACM, 2019), with approximately 1.3 million passenger trips and 148 million ton kilometers of freight transport every day. Under normal conditions, trains arrive and depart according to the timetable and only minor variations in the train service are observed. These variations, which result from differences in driving behavior and from the fact that processes such as coupling and converting may take longer than specified, are referred to as disturbances (Cacchiani et al., 2014). Larger variations, which involve an unexpected change caused by the failure of infrastructure, breakdown of vehicles, unscheduled maintenance, extreme weather or other external events, are referred to as disruptions (Bešinović, 2020). A disruption in the railway network generally means that certain trains experience serious delays or are not capable of running at all (Simons, 2019). Where disturbances are handled by adjusting only the timetable, disruptions require additional adjustments to the rolling stock and crew planning (Mattsson & Jenelius, 2015; Zilko et al., 2016). Recovery models and algorithms have been designed for this purpose, see for example Veelenturf (2014).

Given the complex and dynamic nature of railway networks, it is acknowledged that disruptions will inevitably occur (Zilko et al., 2016). When they do, the emphasis should be on minimizing the consequences and recovering from the impact (Schipper & Gerrits, 2018). Because of the network's intrinsic characteristics, disruptions easily propagate through the network in time and space (Cats & Jenelius, 2014; Malandri et al., 2018). This is partly due to the relative scarcity of infrastructure (Büchel et al., 2020) that makes railways and other public transport networks more sensitive to disruptions compared to road networks (Mattsson & Jenelius, 2015). Delay propagation may even build up to a system-wide scale, leading to new constraints elsewhere in the network (Dekker & Panja, 2021). When this is the case, primary delays will have caused extensive secondary delays, which are defined as delays due to a path conflict between trains (Goverde & Hansen, 2013). A system-based perspective is therefore essential to evaluate and understand the performance of the network during disruptions.

Considering the growing transport demand, a corresponding increase in maintenance works (Van Aken et al., 2017) and the effects of climate change, the number of railway disruptions is expected to increase in the future (Bešinović, 2020). This trend can already be seen building up over the past decade. Since the introduction of a new measurement system by the Dutch Railways (*Nederlandse Spoorwegen, NS*) in 2017, an average of 14 disruptions have been reported daily (Rijden de Treinen, n.d.). Through the years, the numbers have been ever increasing except for 2020, when a 49% decrease in public transport use was observed compared to the year before (Metselaar, 2021). This could be explained by the disrupting effects of the COVID-19 pandemic, which makes data for 2020 and 2021 confounding. Regardless, the overall trend seems to be increasing, as the number of disruptions in 2020 was higher than in 2017. This holds for the absolute number of disruptions as well as for the number of disruptions per train kilometer. Figure 1.1 shows this development over time. The upward trend is problematic as it increases the risk of nonperformance in the medium to long term. In the short term, the

upcoming frequency increase on the RoSA corridor, which will offer one intercity train every ten minutes between Rotterdam and Arnhem via Schiphol, already raises concerns among traffic controllers about potential out-of-control situations in case of future disruptions*.

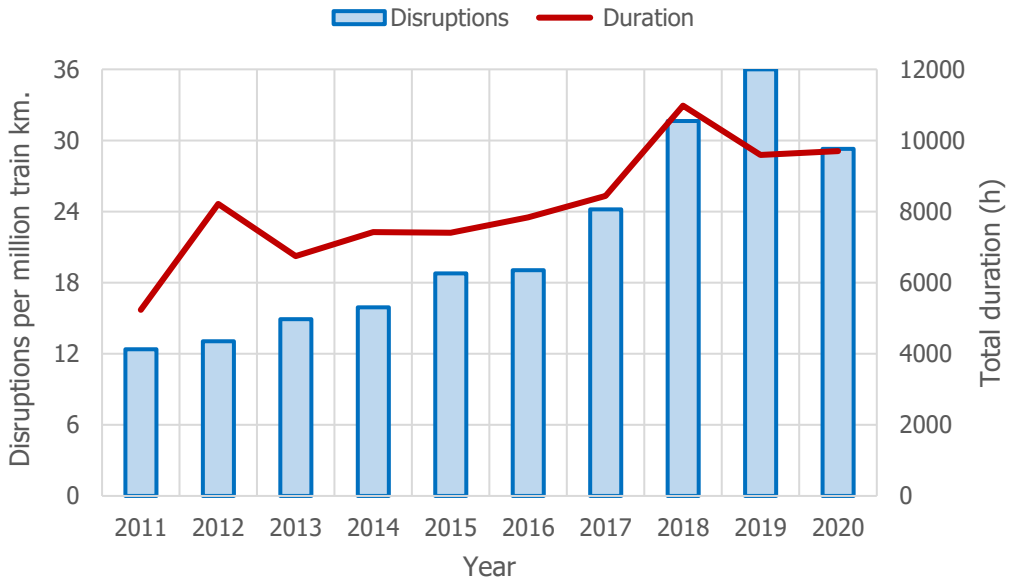


Figure 1.1. Disruptions in the Dutch railway network over time; data from Rijden de Treinen (n.d.).

Company context

Railway disruption management means dealing with the impact of a disruption "to ensure the best possible service for passengers" (Jespersen-Groth et al., 2009) by adjusting the timetable, rolling stock and crew planning. The way in which this is organized is different in each country (Schipper & Gerrits, 2018). In the Netherlands, state-owned infrastructure manager ProRail holds a central position in the disruption management process. ProRail came into existence from the reorganization of the NS in 1995 due to European legislation which forced the exploitation and management of railway infrastructure to be separated. The three subsidiaries that resulted from this reorganization started collaborating in 2003 under the trademark ProRail as part of Railinfratrust BV, the legal owner of most of the Dutch railway network. In 2005, the subsidiaries merged into ProRail BV.

As the infrastructure manager, ProRail has many responsibilities. It is in charge of construction and maintenance of railway tracks; maintenance of train stations; monitoring and coordination of railway traffic; information supply towards train operating companies; and the allocation of network capacity. Disruptions are managed in cooperation with the actors involved and are coordinated nationally from the Operational Control Center Rail (OCCR) in Utrecht. Disruption management is an important part of railway traffic control, which aims to mitigate the consequences of a disruption and keep the traffic outside the disrupted area run smoothly. However, disruption management has to continue evolving if railway system performance is to be kept at the current level. That is where this thesis aims to contribute.

1.2. Problem statement

The research problem addressed in this thesis can be traced back to the limited quantitative knowledge in scientific literature about disruptions in railway networks. In general, not much is known about system behavior during disruptions (Dekker & Panja, 2021), which is typically illustrated with the bathtub model in Figure 1.2 (Ghaemi et al., 2017). The model is more commonly referred to as the resilience curve, because it shows how the system eventually recovers from a disruption. Three phases are distinguished in the curve: the first and third phase are

transition phases, whereas the second phase represents disrupted but stable system behavior. Until now, the resilience curve has remained mostly theoretical in nature, which is why this 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 well understood. Because of this limited quantitative knowledge, designing appropriate measures for disruption management is challenging (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 knowledge would contribute to the effective allocation of resources to prevent, mitigate and recover from disruptions (Malandri et al., 2018).

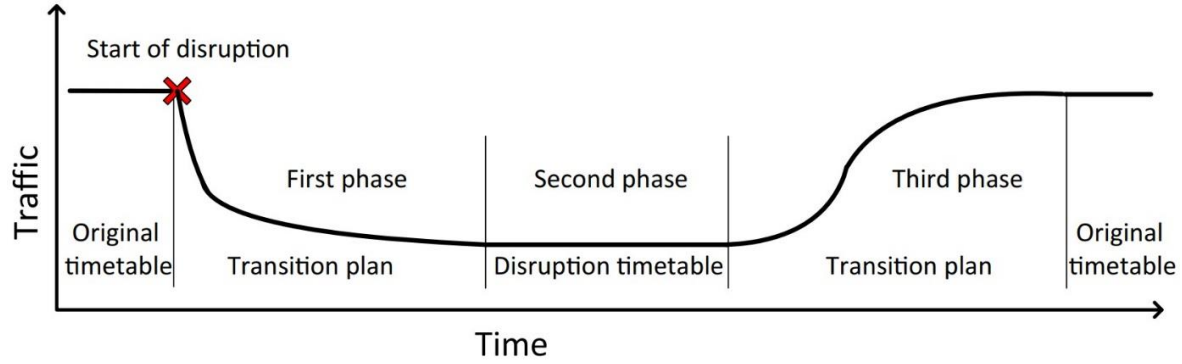


Figure 1.2. The bathtub model (Ghaemi et al., 2017).

The consequences of the current problem are experienced in the first place by railway traffic controllers. They are provided with predefined solutions known as contingency plans in order to manage traffic during a disruption. The goal of a contingency plan is to reach steady state behavior by following an adjusted timetable until the regular service can be restarted. Although this allows quick implementation, the static plans do not account for the dynamic operational environment. For example, the train composition or available infrastructure may be different than anticipated, or a train may have already passed the location where it was meant to short-turn. Consequently, small deviations can render the plans infeasible (Schipper & Gerrits, 2018) and a suitable plan may not always be available or applicable (Ghaemi et al., 2017). Furthermore, the plans do not offer support during the transition phases (Ghaemi et al., 2018). The effects of this problem are felt by train passengers in the form of cancellations and delays. In addition, disruptions in transport systems may lead to economic losses (Tsuchiya et al., 2007). Disruptions in the Dutch railway network can affect logistics processes in the ports of Antwerp and Rotterdam, which has consequences for the hinterland. They can also disrupt international passenger traffic or cause temporary production stops in factories with just-in-time assembly lines*. The presence of these indirect effects adds to the magnitude of the problem.

1.3. Objectives and approach

The main objective of this research is to gain a better understanding of the evolution of railway system performance during disruptions, particularly regarding the dynamics between impact and recovery for the different resilience phases. From a scientific point of view, the objectives are to develop a resilience evaluation approach for analyzing these dynamics and to explicitly describe the differences and similarities between disruptions by empirical testing of the approach, thereby enhancing the quantitative knowledge of resilience. From a societal point of view, the objective is to improve current disruption management practices by contributing to the design and application of recovery measures, with the idea that better quantitative knowledge will lead to more clear design requirements.

The approach to achieve these objectives is to develop a data-driven evaluation approach to make an ex post assessment of the resilience of railway networks using historical traffic realization data and disruption data. In the resilience evaluation, the resilience curve is reconstructed for a large and heterogeneous set of disruptions. Therefore, it is first necessary to determine which performance indicators are most suited to measure railway system performance and which metrics can be used to quantify the shape of the resilience curve. The values of the resilience metrics are evaluated for all disruptions simultaneously in statistical analyses to identify differences and similarities among disruptions of different causes. In comparison with other types of approaches (e.g. topological, optimization-based, simulation-based), the benefits of this data-driven approach are that it removes the need to model the traffic conditions in the network explicitly, and also, that it allows a direct comparison between what practitioners believe to be true and what happens in reality. The downside of a data-driven approach however is that much depends on the availability of sufficient, complete and good-quality data.

1.4. Research questions

Based on the objectives, the main research question is as follows.

Main research question: How does the system performance of a railway network develop during disruptions?

The main question has been divided into five subquestions which represent the theoretical and practical aspects of the research design. Each of Chapters 2 to 6 concludes with the answer to one of the subquestions, which are the following:

1. What can be learned from previous quantitative, data-driven approaches for resilience evaluation of railway networks?
2. What is the current state of the practice and quantitative knowledge regarding different types of railway disruptions in the Netherlands?
3. How can the spatiotemporal effects of disruptions and recovery measures on railway system performance be quantified for the different resilience phases?
4. Which approach should be taken to evaluate railway network resilience for a large and heterogeneous set of disruptions based on traffic realization data?
5. Which insights do quantitative differences and relationships between the resilience metrics bring that may help practitioners evaluate and improve the quality of railway disruption management?

1.5. Research methods

The research questions are answered by applying a combination of methods including a literature review, interviews and quantitative data analysis. The methods are discussed one by one.

Literature review

The first method is a literature review on the quantification of resilience in railway networks. A broader perspective was taken by also considering air, subway, waterway, freight and general (public) transport systems. Supply chain resilience was included as an alternative domain. The motivation behind this choice is given at the start of Chapter 2. The review discusses resilience definitions, performance indicators, evaluation approaches and resilience metrics. It also highlights the research gaps that are addressed in this thesis. Scopus was used as the main source for finding articles based on a keyword search within the article title, abstract and keywords. Scientific papers as well as conference papers were collected. After completing the keyword search, backward snowballing (Jalali & Wohlin, 2012) was applied to find additional articles, with an emphasis on the references included in previous literature reviews on resilience.

Interviews

The second method is a series of open, semi-structured interviews with practitioners. This helps give direction to the research in the beginning, discover what existing knowledge there is to build on and understand the story "behind the data". Interviews were conducted throughout the course of the research. The scope and purpose of the interviews changed as the research progressed. Table 1.1 presents the scope prior to and during data collection, during the experiments and during data evaluation. An overview of the respondents is included in Appendix C. Relevant findings from the interviews have been incorporated throughout the report. In-text references are marked with an asterisk (*), keeping the references anonymous.

Table 1.1. Scope of the interviews per research phase.

Data collection	Experiments	Data evaluation
General working practices	Interpretation of the data	Recognizability of the results
Design and application of contingency plans	Limitations of the used methodology	Suggestions improvements for disruption management
Insights from earlier projects	Relation between theory and practice	Effect of future developments on the disruption dynamics
Performance measurement		Contributions in the short term
Data structure and availability		
Peculiarities in the data		
Potential hypotheses		

Data analysis

Building on the theoretical and practical insights from the literature review and interviews, the third and final method is a quantitative data analysis. The analyzed data concerns historical railway traffic realization data and disruption log data. Realization data was used to calculate system performance during disruptions based on the time and location obtained from the disruption log data. Disruptions were categorized into groups and their resilience metrics were compared in group comparisons. Several options are available for group comparisons, which include one-way analysis of variance (ANOVA), the Kruskal-Wallis test and Welch's ANOVA. Additional relationships between the metrics were explored in regression analysis. The data analysis is organized according to the general procedure shown in Figure 1.3.

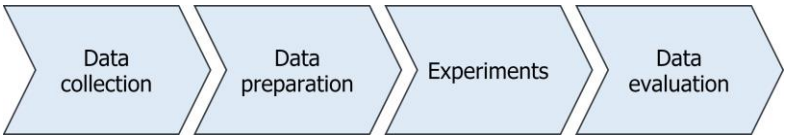


Figure 1.3. General data analysis steps.

Data collection: Data was collected from ProRail's data analysis application "Sherlock", which brings together data from multiple sources in four categories: train movements, infrastructure, plan mutations and disruptions. As the application was not designed to export large amounts of data, the realization data was requested directly from the involved department. Appendix D contains a data collection summary that specifies which data was collected, when, by whom, and in which version of Sherlock.

Data preparation: Data preparation involves the structuring, cleaning and processing of the collected data prior to analysis. More specifically, this concerns handling invalid, inconsistent and missing data entries; correctly interpreting the part of the data in Sherlock that have been preprocessed or approximated; and aggregating the realization data in time and space. Data preparation is discussed in more detail in Chapters 4 and 5.

Experiments: The experiments broadly consist of studying an example case, identifying representative resilience curves and performing statistical analyses. Prior to analysis, it is necessary to check if the assumptions of the chosen statistical test are satisfied. If assumptions are violated, a test cannot be used to draw valid conclusions. The results of a test are referred to as the test statistics, which indicate whether or not the result is significant. Group comparisons such as ANOVA do not reveal exactly where the differences are; they only indicate whether at least one of the group means or medians is significantly different from the others (Mertens et al., 2017). To gain insight into the actual differences, post hoc tests are required. For ANOVA and Kruskal-Wallis, a range of post hoc tests is available. Statistical methods are discussed in more detail in Chapter 4.

Data evaluation: Data evaluation involves the verification of the results and the identification of factors that may have contributed to the results, which is necessary before drawing any conclusions. Data evaluation has been integrated in the storyline of the experiments in this report and is discussed in Chapter 5.

All experiments were coded in Python using Jupyter notebooks. Relevant libraries and modules include Pandas (McKinney, 2007), Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021), Scipy, Researchpy, Pingouin (Vallat, 2018) and Statsmodels (Seabold & Perktold, 2010).

1.6. Research scope

This thesis research was carried out at ProRail with the support of the departments VGB, PAB and A&E (see List of abbreviations). The scope of the research concerns the evolution of system performance during disruptions in the Dutch railway network in 2019. As this comes down to determining the exact shape of the bathtub, only disruptions with an observable impact on the train service were included. In Sherlock, these are events for which a logistical record exists, in which case a capacity reallocation and usually also a contingency plan have been applied. The logistical record contains the logging of an event. In the remainder of this report, those events are simply referred to as disruptions. Why or at what interval a certain disruption occurred is irrelevant for this research. The fact that a disruption occurred is taken for granted; the focus is on what happened as a result. In certain cases, multiple disruptions may occur around the same time and in the same geographical area. Those disruptions were ultimately considered out of scope. Also, it should be stressed that the resilience of the railway network is described in terms of railway traffic, and not in terms of the physical infrastructure or individual infrastructure elements. Lastly, the spatial propagation and potential cascading of disruption effects (in the form of cancellations and delays) were incorporated in the resilience curve by the way in which performance was aggregated. An alternative approach would have been to perform what is referred to in ProRail as an “oil stain analysis”. This involves analyzing which trains experienced secondary delays due to a disturbance or disruption elsewhere. Although this could be done in Sherlock, it would require a different setup of the research design that focuses more on the individual train level.

In short, this research **does not** cover the following:

- Events where no measures were applied
- Causal factors leading to a disruption
- Interdependency of disruptions over time
- Disruptions with an overlapping area or time period
- Resilience of the railway infrastructure or individual infrastructure elements
- “Oil stain analysis” of secondary delays

1.7. Scientific and societal significance

This research, which started from a scientific perspective, has both scientific and societal relevance. The main scientific contribution consists of the new data-driven approach, which is structured in a resilience evaluation framework, to ex post assess the resilience of railway networks. The approach is generic and could be applied to railway networks in other countries, or potentially, other types of (transport) systems as well. Additional contributions include a composite performance indicator capable of representing delays and cancellations simultaneously, two new resilience metrics to quantify performance in the transition phases, and a method for identifying the resilience phases from the curve itself. A last scientific contribution consists of empirical testing of the proposed evaluation approach on a case study, thereby improving the quantitative understanding of resilience which at present is still immature.

The main societal contribution consists of recommendations for improving data processing and disruption management practices in ProRail. According to the *Koers van VL*, disruption management should become more standardized and predefined than it is now, which means that predefined solutions should be made for the transition phases as well, and measures need to become more network-oriented. The recommendations made in this report could support that transition. The acquired knowledge about disruption dynamics in the Dutch railway network is a second contribution, which could ultimately increase the predictability towards passengers during disruptions. A third contribution is the potential this research has demonstrated for future evaluation studies involving traffic realization data, by showing that the necessary data are available and that relevant conclusions can be drawn from the analysis of large numbers of disruptions.

1.8. Reading guide

The remainder of this report is structured as follows. Chapter 2 presents the literature review on resilience definitions, performance indicators, evaluation approaches and resilience metrics. It also identifies the existing research gaps. Chapter 3 discusses disruption management working practices, first generally and then specifically for the Dutch case. Chapter 4 explains the methodology according to the resilience evaluation framework. Chapter 5 introduces the case study and presents the results of the experiments. Chapter 6 discusses the results in depth and reflects on the scientific and practical implications. Chapter 7 presents the conclusions, recommendations, limitations and future research directions. The structure of the report is summarized in a flowchart in Figure 1.4.

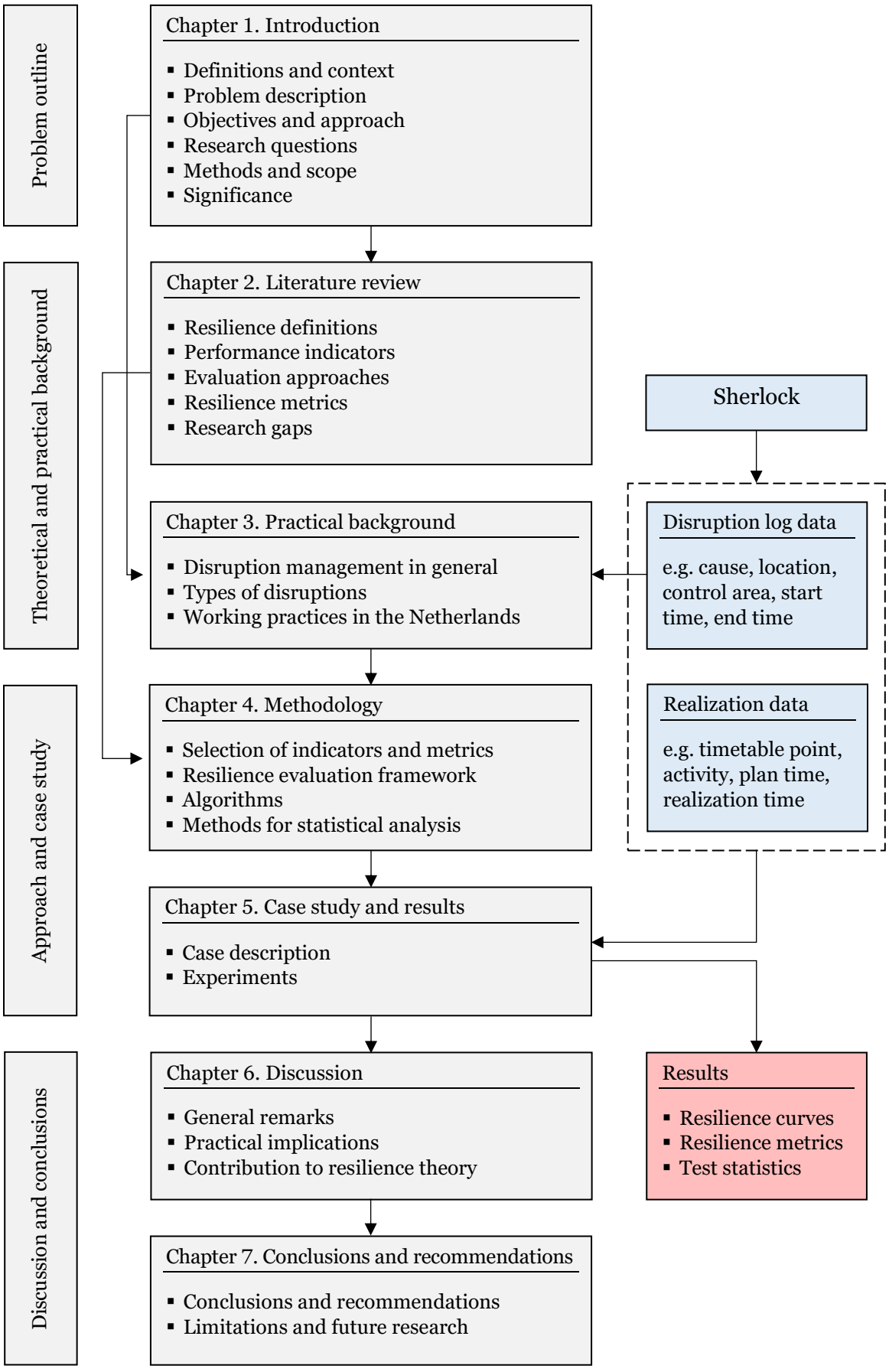


Figure 1.4. Flowchart of the report structure.

2. Literature review

This chapter provides a literature review of the different theories and research avenues related to the research problem that is addressed in this thesis. The review is structured according to Van Wee and Banister (2016). Thus, Section 2.1 introduces the scope and purpose, addresses previous reviews and gives a general overview of resilience. Sections 2.2 to 2.5 form the main body of the review. Section 2.2 explores various resilience definitions and presents a revisited definition to be used throughout this report. Section 2.3 explores performance indicators that can be used to describe railway system performance. Section 2.4 discusses previous data-driven resilience evaluation approaches. Section 2.5 gives an overview of resilience metrics that can be used to describe the profile of the resilience curve. Section 2.6 identifies the existing research gaps in railway resilience literature. Finally, Section 2.7 summarizes the chapter and provides the answer to subquestion one.

2.1. Scope and purpose

Resilience is not just a relevant topic in railway transport, but in other research domains as well. Because resilience research exhibits domain-independent characteristics (Madni et al., 2020), the scope of this review is not limited to railway related studies. While the emphasis is on resilience in railway transport, the review also covers air, subway, waterway, freight and general (public) transport systems. Where applicable, references are included to articles that adopt a general systems view. Additionally, supply chain resilience is covered as an alternative perspective, since a comparable problem of limited quantitative knowledge exists in the supply chain domain (Ivanov et al., 2014) and references to supply chain resilience appear occasionally in earlier transport related work, such as Gonçalves and Ribeiro (2020). Also, there is currently a growing interest in this topic due to the disrupting effects of COVID-19 (Van Hoek, 2020). Combining these research domains could lead to new insights into the quantification of resilience. Moreover, it is acknowledged that resilience is a fundamentally different construct than robustness (Brandon-Jones et al., 2014) and vulnerability (Mattsson & Jenelius, 2015), which are therefore not addressed. The difference between resilience and robustness is briefly discussed in Section 2.2. The purpose of this review is to assist in answering subquestions one and three. This requires formulating a definition of resilience to build on; identifying ways to define railway system performance; studying what can be learned from previous data-driven approaches; and identifying which resilience metrics are available to choose from.

Several review papers have been published in the past addressing the resilience of general systems (Hosseini et al., 2016), transport systems (Mattsson & Jenelius, 2015; Zhou et al., 2019; Gonçalves & Ribeiro, 2020), railways (Bešinović, 2020) and supply chains (Tukamuhabwa et al., 2015; Pettit et al., 2019). Also, Cacchiani et al. (2014) reviewed rescheduling models and algorithms for railway disruption management. Of particular interest for the review in this chapter are the resilience definitions provided in the various reviews; the discussion of metrics by Hosseini et al. (2016) and Zhou et al. (2019); and the classification of resilience evaluation approaches for railway systems by Bešinović (2020). Models and algorithms for disruption management are not addressed in this chapter because they do not help with the quantification of system performance. Besides, reviewing such models could easily fill an entire new chapter.

The previous review papers help create a general overview of resilience. It appears that a range of resilience definitions have been proposed across research domains. Both qualitative and quantitative resilience metrics have been proposed (Hosseini et al., 2016) which are used to describe system performance during disruptions (Gonçalves & Ribeiro, 2020). Disruptions can have either internal or external causes (Mattsson & Jenelius, 2015). The choice of indicators and metrics depends on the evaluation approach (Zhou et al., 2019). For example, topological indicators usually account for the topological structure of the network, whereas performance-

based indicators may put more focus on the traffic flow. Quantitative evaluation approaches in railway transport include topological approaches (e.g. Dorbritz, 2011), which assess network resilience based on the topological structure; optimization approaches (e.g. Azad et al., 2016; Van Aken et al., 2017), which apply mathematical models for improving resilience; simulation approaches (e.g. D'Lima & Medda, 2015; Adjetey-Bahun et al., 2016), which model the impact and response in hypothetical scenarios; and data-driven approaches (e.g. Chan & Schofer, 2016; Janić, 2018), which evaluate system performance by studying historical data. As this thesis adopts a data-driven approach, mainly articles in the last category were reviewed.

2.2. Definitions of resilience

It is acknowledged that there is no unique way of defining resilience in literature (Hosseini et al., 2016). Several theoretical definitions are therefore compared in this section. The system states that occur before and during a disruption are explained as well. To conclude, a revisited definition of the resilience of transport systems is provided.

Theoretical definitions

To gain a proper understanding of resilience, definitions from multiple research domains are studied in this review. Reviews of resilience definitions have been performed in the past, for example by Zhou et al. (2019) and Gonçalves and Ribeiro (2020) in transport and by Tukamuhabwa et al. (2015) in supply chains. Some definitions are more narrowed down to a single research domain than others, but all can be generalized to describe the resilience of an arbitrary system. As not all definitions are equally comprehensive, it helps to classify them based on the aspects that they did or did not incorporate.

Two recent articles (Madni et al., 2020; Bešinović, 2020) made such a classification. Madni et al. (2020) found that the resilience of general systems can be defined along one of four dimensions: the capacity to rebound, resist, adapt, or adapt continuously. Following their line of argument, a definition of resilience for systems modeling is only useful if it offers some insight into the implementation while accounting for finite resources and the presence of performance boundaries. This requires viewing resilience as an adaptive capacity. Bešinović (2020) gave a review of resilience definitions in railway resilience literature and found that most articles relate to the "ability to recover quickly from a disruption".

The latter definition may be extended in a number of ways, as shown in Table 2.1. The return to an original or acceptable system state may be assumed implicitly for all definitions, though some articles mention this explicitly. Most articles also specify the limited availability of time, and potentially, other resources. Furthermore, according to a few articles, the system must be able to maintain acceptable performance once the disruption impact has been absorbed. This notion differs from the ability to withstand disruptions, which should instead be referred to as robustness (Madni et al., 2020).

Table 2.1 presents a classification of resilience definitions. Some of the definitions referred to in the table make the mistake of creating an equivalence between resilience and robustness. This difference is often not properly understood (Brandon-Jones et al., 2014), which is especially confusing in the railway context, where robustness refers to the ability to withstand variations in daily operation (Bešinović, 2020). Notice the similarity with the definition from Madni et al. (2020). Still, arguing that resilience should only describe the events that take place in the aftermath of a disruption would be flawed, as it cannot be denied that there is a proactive side to resilience which helps prevent disruptions (Bešinović, 2020) and reduce their impacts. For this reason, Melnyk et al. (2014) divided resilience into two capacities: resistance and recovery. This again emphasizes the importance of a consistent terminology. For resilience, a revisited definition is given at the end of this section. For robustness, the definition commonly used in railway transport is maintained.

Table 2.1. Theoretical definitions of resilience in previous studies.

Reference	Ability to recover from and/or withstand disruptions	...returning to the original state	...within a reasonable amount of time	...while maintaining acceptable performance
Bešinović (2020)	x		x	
Brandon-Jones et al. (2014)	x	x	x	
Bruneau et al. (2003)	x		x	
Carvalho & Cruz-Machado (2011)	x			
Cats & Jenelius (2014)	x		x	
Chan & Schofer (2016)	x	x	x	
Chen & Miller-Hooks (2012)	x			
D'Lima & Medda (2015)	x	x	x	
Gonçalves & Ribeiro (2020)	x	x	x	x
Janić (2015)	x			
Jin et al. (2014)	x			x
Ponomarov & Holcomb (2009)	x			x
Ren et al. (2020)	x	x	x	
Rodrigue (2020)	x	x	x	

System states

As a general construct, resilience is thought to comprise a number of consecutive system states in which the system performs differently. These states represent phases in the resilience curve as shown in Figure 2.1. The curve is also known as the bathtub model (Ghaemi et al., 2017) because of its shape. The bathtub model may be used as a conceptual characterization of the disruption management process. To be consistent with the nomenclature of the phases in the bathtub model, the first phase of the resilience curve is named “phase 0” so the remaining ones can be named 1 to 3. The word “phase” is used here to emphasize the time dependency, and the word “state” is used to describe the condition of the system during a certain phase. Table 2.2 shows how researchers have interpreted the system states and which definitions are used in this thesis. The findings were united in the following terminology:

- **Phase 0: Preparedness.** Indicates the normal operation of the system with the knowledge that a disruption may occur at some point in time. Characterizing this phase as “avoidance” or “prevention” is therefore futile.
- **Phase 1: Degradation.** Indicates the decrease in system performance due to the disruption. This phase involves survival of the initial impact and, if possible, application of emergency measures to reduce the impact until the system has been stabilized.
- **Phase 2: Response.** Indicates the new, disrupted steady state of the system. This phase involves assessment of the initial damage caused by the disruption and application of temporary measures, such as contingency plans in railways, to prepare for recovery.
- **Phase 3: Recovery.** Indicates the return to an acceptable performance level, resolving the effects of the disruption. Recovery may be full or only partial. In railway transport, recovery means reinserting the canceled services and returning to the original timetable (Ghaemi et al., 2017).

Table 2.2. Definitions of system states in previous studies.

Reference	Phase 0	Phase 1	Phase 2	Phase 3
Altay & Green (2006)	Preparedness	Response	-	Recovery
Baroud et al. (2014)	Reliability	Survivability	-	Recoverability
Bešinović (2020)	Robustness	Survivability	Response	Recovery
Bevilacqua et al. (2017)	Prevention	Mitigation	-	Recovery
Gonçalves & Ribeiro (2020)	Resistance	Absorption	Transformation	Recovery
Melnyk et al. (2014)	Avoidance	Containment	Stabilization	Return
Ouyang et al. (2012)	Prevention	Propagation	Assessment	Recovery
This thesis	Preparedness	Degradation	Response	Recovery

In Figure 2.1, the linearized performance is shown during the four phases of a disruption. Each timepoint T_i marks the end of a phase. The impact is defined as the difference between original and disrupted performance. This depiction assumes that performance degrades and recovers gradually. Depending on the type of disruption, this may not be the case and not all resilience phases may be equally observable. In case of a sudden shock, such as an earthquake or power outage, the decrease in performance is abrupt (e.g. Bruneau et al., 2003) and there would be no degradation phase. Likewise, there may or may not be a steady response phase prior to recovery. For the general case however, the profile shown in Figure 2.1should be assumed.

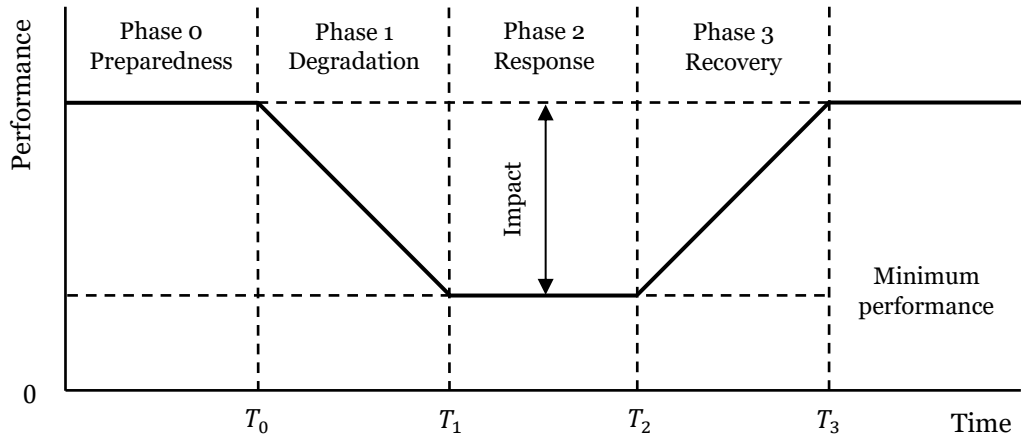


Figure 2.1. Schematic illustration of the resilience curve.

Revisited definition

Based on the prior considerations, an inclusive definition of resilience should represent the different system states. Also, from a customer's (here: passenger's) perspective, it is important that recovery takes place within a reasonable amount of time and without causing too much inconvenience. Maintaining an acceptable level of service in the disrupted state is therefore desirable. This leads to the following revisited definition of resilience, which is mainly a synthesis of Gonçalves and Ribeiro (2020) and Bešinović (2020).

Resilience is the ability of a system to prepare for a disruption, as well as to reduce, absorb and accommodate the impact of a disruption while maintaining an acceptable level of service, and to recover to a desired state of operation within a reasonable amount of time.

2.3. Performance indicators

The performance of a system must be specified first in order to quantify its resilience. For this purpose, a time-dependent performance indicator or figure-of-merit has to be selected (Henry & Ramirez-Marquez, 2012). In a two-dimensional representation such as Figure 2.1, with time on the horizontal axis, this means that the spatial aspect of the loss of performance must somehow be included on the vertical axis. Thus, considering the spatial propagation of disruption effects comes down to selecting the proper performance indicator and aggregating the performance measurement in space. Alternatively, space could be included on the horizontal axis instead of time, or on a third axis in a three-dimensional representation. With that in mind, this section reviews indicators of transport system performance that are suitable for measuring railway system performance.

Four categories of indicators were identified in this review: travel time, travel demand, traffic flow and use of resources. Within these categories, the discussion focuses on indicators that are real-time ready, which means they are able to measure performance at any moment in time. This is key as the resilience curve is usually drawn as a function of time. Indicators that can only be used to measure performance in retrospect over a period of time, such as passenger welfare (Cats & Jenelius, 2014), revenue vehicle miles (Chan & Schofer, 2016) and overall travel time (De-Los-Santos et al., 2012) do not meet this criterion and were therefore not considered. The different real-time performance indicators per category are discussed below.

Travel time

Indicators that represent travel time are *punctuality*, *journey time*, *average delay*, *secondary delay* and *cumulative delay* (e.g. Goverde & Hansen, 2013; Nicholson et al., 2015). Punctuality is expressed in percentages, where the other indicators are expressed in seconds or minutes. When a disruption occurs, punctuality is expected to decrease, but this reveals little information about the size of the impact. It only tells that some trains are delayed, but not by how much. Therefore the average, secondary or cumulative delay may be considered. Alternatively, the average journey time could be used, but the journey time depends on the types of services and the network structure in the disrupted area. Delay indicators are more practical in application than the average journey time since target performance is, obviously, zero.

Travel demand

Indicators that represent travel demand are *satisfied demand* (e.g. Jin et al., 2014), *passenger volume* (e.g. D'Lima & Medda, 2015) and *network saturation* (Malandri et al., 2018). Satisfied demand is expressed in percentages and denotes which part of the demand can still be satisfied during or after a disruption. In contrast, the passenger volume denotes the number of people traveling in the network. Of the two indicators, satisfied demand should be preferred because it allows a direct comparison with a reference value. Network saturation is expressed in percentages and accounts for the redistribution of passengers in the network. A high saturation indicates potential overcrowding on certain routes when people are taking alternative routes, though in reality, this may not always be possible. A downside of the travel demand indicators is that they require passenger counts, or at least accurate estimates of passenger numbers.

Traffic flow

Indicators that represent traffic flow are *traffic throughput* (Ghaemi et al., 2017; Jafino et al., 2020) and *speed reduction* (Gonçalves & Ribeiro, 2020). Traffic throughput indicates the local number of train passings on a certain route or along the cross-section of an area, which will typically be lower than planned in case of a disruption. Speed reduction naturally contributes to trains being delayed, but normal operating speed varies with the train type, type of service and track segment, which makes it hard to bring this indicator into practice.

Use of resources

Indicators that represent the use or availability of resources (e.g. rolling stock, infrastructure) are *track occupation*, *rolling stock usage*, *canceled services* and *transport capacity* (e.g. Nicholson et al., 2015; Janić, 2018). Track occupation and rolling stock usage are expressed in percentages and are assessed by comparing the actual situation to the original plan. As opposed to rolling stock usage, canceled services may denote either the percentage of trains that are not running or the percentage of canceled train activities (e.g. departures, arrivals). Canceled services thus offer a valuable addition to the travel time indicators. Transport capacity denotes the available seats or seat-kilometers, which is more difficult to determine as it depends on the type and composition of the trains that are deployed. In addition, a common problem for these indicators is that (with the exception of canceled services) one cannot say with certainty that the effects of a disruption have dissipated once the indicator values return to normal.

Overview of indicators

Table 2.3 provides an overview of the discussed performance indicators. When the direction of the indicator is positive, such as for punctuality, a higher value suggests better performance. This is consistent with how the resilience curve has been depicted so far. When the direction is negative, such as for delays, a higher value suggests worse performance. This is referred to as a decreasing system service function (Baroud et al., 2014). Between the available indicators, there are differences with regard to ease of measurement and the extent in which they reflect the interests of passengers, train operating companies (TOC) and the infrastructure manager (IM). The passenger's perspective is covered by the travel time and demand indicators, while traffic flow or resource indicators may be equally or even more telling for the IM and TOCs.

Table 2.3. Performance indicators used in previous studies.

Category	Indicator	Direction	References
Travel time	Punctuality	Positive	Evans (2011), Goverde & Hansen (2013), Woodburn (2019)
	Journey time	Negative	Nicholson et al. (2015)
	Average delay	Negative	Adjetey-Bahun et al. (2016), Büchel et al. (2020), Goverde & Hansen (2013)
	Secondary delay	Negative	Goverde & Hansen (2013)
	Cumulative delay	Negative	Evans (2011), Janić (2018), Nicholson et al. (2015)
Travel demand	Satisfied demand	Positive	Chen & Miller-Hooks (2012), Jin et al. (2014)
	Passenger volume	Positive	Adjetey-Bahun et al. (2016), D'Lima & Medda (2015)
	Network saturation	Negative	Malandri et al. (2018)
Traffic flow	Traffic throughput	Positive	Ghaemi et al. (2017), Jafino et al. (2020)
	Speed reduction	Negative	Gonçalves & Ribeiro (2020)
Use of resources	Track occupation	Positive	Goverde & Hansen (2013), Nicholson et al. (2015)
	Rolling stock usage	Positive	Nicholson et al. (2015), Woodburn (2019)
	Transport capacity	Positive	Janić (2018), Nicholson et al. (2015)
	Canceled services	Negative	Evans (2011)

2.4. Evaluation approaches

As mentioned in Section 2.1, the data-driven approach is one of four possible types of resilience evaluation approaches. Data-driven approaches are promising for the ex post evaluation of disruptions considering the growing demand for the quantification of system performance during disruptions, and thus, resilience (Bešinović, 2020). As this thesis follows such a data-driven approach, a look into previous data-driven evaluation approaches is considered valuable. Data-driven approaches in railway transport, other public transport (subway and taxi) and air transport are discussed. Because no comparable approaches were found in the supply chain domain, simulation approaches are discussed for supply chain resilience evaluation instead.

Resilience evaluation in railways

Several approaches for the resilience evaluation of railway networks are discussed. Firstly, Chan and Schofer (2016) studied the resilience of a railway network subject to extreme weather events. The recovery of the New York City heavy rail transit was examined after the blizzard of 2010 and hurricanes Irene and Sandy based on lost revenue vehicle miles, restoration time and lost service days. Janić (2018) fitted an analytical model to historical data from the Japanese high-speed rail network after the 2011 earthquake. He linearized the resilience curve similar to Figure 2.1 and calculated resilience for each phase as the ratio of actual to planned performance. Aggregate values of network resilience were obtained by summing over the lines, routes and indicators. Three indicators were specified for each of eight types of performance. In short, one could say that multiple resilience curves were combined into one curve and an aggregate metric was derived, instead of deriving multiple metrics from a single curve.

Woodburn (2019) studied the consequences of a lengthy, unplanned closure of a major railway freight route in Britain. Open access data collected at the individual train level was used to demonstrate a gradual improvement in traffic and service levels in the long term, although the discussion of the results leaves room for interpretation. Büchel et al. (2020) studied delay propagation in the Swiss railway network after a large-scale disruption in Germany by comparing arrival delays for a disrupted and undisrupted scenario. They found that during the disruption, significantly smaller delays were experienced due to the lower variability in operations. The scenario was replicated in a simulation, which is out of scope for this review.

Resilience evaluation in other public transport modes

Looking at other public transport modes, Zhu et al. (2016) studied hurricanes Irene and Sandy using taxi and subway ridership data. The recovery curves, expressed in terms of the recovery rate for each evacuation zone, reveal a different recovery behavior for the two events and between zones. Spatial dependence of the resilience per zone was further investigated by Zhu et al. (2017) by presenting the recovery rate, recovery time and loss of resilience in geospatial displays such as in Evans (2011). Similar displays are also found in Malandri et al. (2018) and Büchel et al. (2020). Ren et al. (2020) constructed a Bayesian network by evaluating data for approximately 50,000 disruptive events in the Beijing subway between 2013 and 2018. They demonstrated the causal relationships between what they referred to as the fault case, failure mode and influence mode. Although this study did not involve performance measurement, it stands out due to the large amount of collected data.

Resilience evaluation in air transport

Another mode of transport subject to resilience research is air transport. The resilience and friability (i.e. the decrease of resilience due to the removal of network components) of the air transport network around New York LaGuardia was studied by Janić (2015) for a large-scale disruption. An aggregate resilience measure was proposed which considers the duration of the impact, but not the recovery phase. Wong et al. (2020) followed a hybrid data-driven network analysis approach to search for abnormalities in arrival delays by using a specific statistical measure called the Mahalanobis distance. Instead of observing the larger network, four US airlines were assessed individually, which means that limited insight can be obtained from this study when it comes to network resilience.

Resilience evaluation in supply chains

For supply chains, the lack of quantitative knowledge about resilience is ascribed to the difficulty of collecting data, since supply chain disruptions are generally less observable and more easily confounded by human factors (Macdonald et al., 2018) than those occurring in transport systems. This means that simulation approaches offer a viable alternative. Furthermore, rather than a system property, resilience in supply chain management can be regarded as a paradigm

for organizing supply chain activities (Carvalho & Cruz-Machado, 2011). Most of the effort goes out to finding strategies that improve resilience (Tukamuhabwa et al., 2015), while resilience is rarely analyzed in an operational context. Only a few relevant examples of resilience evaluation approaches in supply chains were found.

Spiegler et al (2012) investigated the relation between resilience and lead time by simulating different inventory and production control strategies, which revealed a tradeoff between resilience and robustness. Munoz and Dunbar (2015) studied the transient response across supply chain tiers in a simulation model. They proposed five resilience metrics to be combined into an aggregate measure of resilience by means of structural equation modeling. Statistical testing showed the importance of two metrics representing nonlinearity, thereby challenging a linear recovery profile. Macdonald et al. (2018) also adopted a simulation approach for the purpose of theory building and investigated through structured experiments how different levels of buffer stock, connectivity and shock interarrival time affect system performance. Performance was defined in terms of four dependent variables. Regression analysis was performed to evaluate the effects, which proved to be significant in most cases.

Overview of approaches

An overview of the discussed approaches is presented in Table 2.4. It is observed that most transport related studies examined a single, large-scale disruptive event. Ren et al. (2020) and Wong et al. (2020) studied a large number of disruptions, but they did not do so to assess the evolution of system performance. Regarding the methods, statistical analyses were frequently used to assess the significance and/or dependence of resilience measures.

Table 2.4. Resilience evaluation approaches followed in previous studies.

Reference	Domain	Approach	Research topic
Chan & Schofer (2016)	Railway	Data-driven	Resilience of a metropolitan railway network in extreme weather
Janić (2018)	Railway	Data-driven	Resilience of a high-speed railway network affected by an earthquake
Woodburn (2019)	Railway	Data-driven	Impacts from the unplanned closure of a railway freight route
Büchel et al. (2020)	Railway	Data-driven, simulation	Delay propagation after a large-scale railway disruption
Zhu et al. (2016)	Subway, taxi	Data-driven	Post-hurricane recovery of taxi and subway trips
Zhu et al. (2017)	Subway, taxi	Data-driven	Spatial modeling of post-hurricane recovery
Ren et al. (2020)	Subway	Data-driven	Causal relationships in subway disruptions
Janić (2015)	Air transport	Data-driven	Resilience of an air transport network during a hurricane
Wong et al. (2020)	Air transport	Data-driven	Ability of airlines to prevent abnormally large delays
Spiegler et al. (2012)	Supply chain	Simulation	Relationship between resilience and lead time
Munoz & Dunbar (2015)	Supply chain	Simulation	Transient response across supply chain tiers
Macdonald et al. (2018)	Supply chain	Simulation	Impacts from supply shocks on system performance

2.5. Resilience metrics

Resilience metrics describe the profile of the resilience curve. Due to the heterogeneous characteristics of disruptions and recovery measures that may be taken, resilience is considered to be a multidimensional construct, and therefore, it is incapable of being captured in a single metric (Munoz & Dunbar, 2015). If only one metric were to be used, then entirely different loss and recovery behaviors could result in the same resilience value (Zobel, 2011). Regarding the type of metrics, Hosseini et al. (2016) distinguished between deterministic and probabilistic system

resilience metrics. As it is assumed that metrics can be derived analytically from the resilience curve, this review considers deterministic metrics only. This section first covers the range of metrics found in railway resilience literature, followed by additional metrics found in supply chain resilience literature.

Metrics in railway literature

Several articles (e.g. Nicholson et al., 2015; Chan & Schofer, 2016) describe general resilience metrics in a railway context. Such general metrics can be defined regardless of the underlying system structure and are therefore comparable across research domains (Hosseini et al., 2016). This explains why all metrics described in this subsection also appear in general systems literature as can be seen in Table 2.5. Metrics that are defined along the time axis are *recovery time*, *recovery rate* and *deterioration rate* (e.g. Chan & Schofer, 2016; Janić, 2018). Recovery time is in fact the most common resilience metric in transport literature (Zhou et al., 2019). Metrics that are defined along the performance axis are *initial impact*, *maximum impact* and *minimum performance* (e.g. Dorbritz, 2011; Nicholson et al., 2015). Minimum performance is also referred to as *residual functionality* (Cimellaro et al., 2010) because it amounts to the original performance minus the maximum impact.

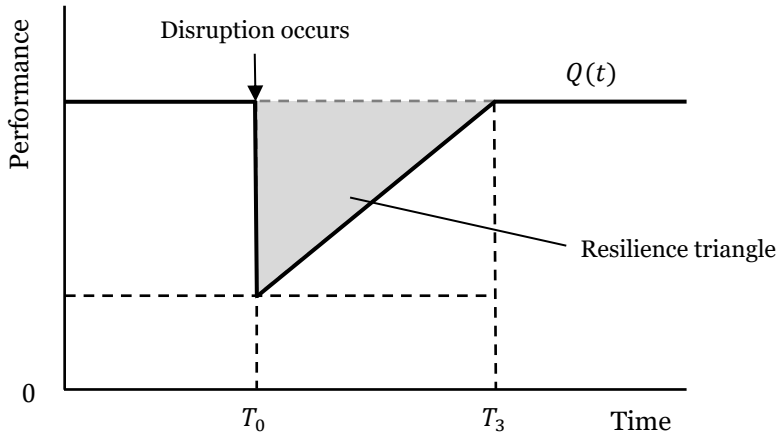


Figure 2.2. Simplified resilience curve and the resilience triangle.

The last metric found in railway literature was originally proposed by Bruneau et al. (2003) for general systems as the *loss of resilience*. Equation (1) shows the loss of resilience R with performance Q at time t expressed in percentages. The integral corresponds to the area of the shaded resilience triangle shown in Figure 2.2, assuming a sudden drop in performance and a linear recovery profile. The figure can be obtained from Figure 2.1 by omitting the degradation and response phases. The loss of resilience may also be referred to as *deviation area* (Nicholson et al., 2015) or *service loss* (Chan & Schofer, 2016). However, in this thesis it is referred to as *performance loss* in accordance with Munoz and Dunbar (2015), for that is what it represents: the cumulative loss of performance over time.

$$R = \int_{T_0}^{T_3} \big(100 - Q(t)\big)\,\mathrm{d}t \tag{1}$$

Additional metrics

Additional metrics are found in supply chain literature, which could be used for the resilience evaluation of railways or other transport systems as well. Based on Equation (1), Spiegler et al. (2012) adopted the *integral of time absolute error (ITAE)* as applied in control engineering, which penalizes a slower return to the steady state. Munoz and Dunbar (2015) adopted two additional metrics that describe the nonlinearity of recovery. The *profile length* equals the

length of the recovery profile, and the *weighted sum* equals the time-dependent deviation from a linear recovery profile. Analytical expressions for all metrics are included in Table 2.6.

Overview of metrics

To summarize, ten resilience metrics are identified which can be allocated to the dimension of time, performance or both. An overview is given in Table 2.5. Metrics that appear differently in the reviewed articles but measure the same value, such as the loss of resilience and performance loss, are included under the same heading. Recovery time is the most commonly used metric. In most of the articles, it describes the time between the moment performance drops below a certain threshold and the moment performance is restored. The second-most common metric is performance loss, which equals the area enclosed by the resilience curve and target performance. The third-most common metric is the maximum impact.

Table 2.5. Resilience metrics used in previous studies.

Dimension	Resilience metric	Domain	References
Time	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)
Performance	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)
Time and performance	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)

Table 2.6 presents the analytical expressions for the identified resilience metrics. Traditionally, $Q(t)$ is used to denote the performance function. Here, $g(t)$ represents the linearized recovery profile. The other symbols are explained as follows:

- T_i is the moment when phase i ends, in accordance with Figure 2.1
- Q_0 represents target performance
- Q_{min} represents minimum performance
- a_k is the time in the recovery phase at interval k

The one-dimensional metrics, included in the left half of Table 2.6, are easily determined from the resilience curve. The two-dimensional metrics however, included in the right half of the table, are defined by slightly more complex analytical expressions. This may require the use of numerical integration methods such as Simpson's rule, even when the performance function is known (Munoz & Dunbar, 2015). Furthermore, there is an overlap between metrics such as recovery time and recovery rate, or maximum impact and residual functionality. Including both such metrics would introduce multicollinearity, which means that the value of one metric could be derived directly from the other metric with relative certainty. Multicollinearity should be avoided in data analysis (Mertens et al., 2017).

Table 2.6. Analytical expressions for the identified resilience metrics.

Resilience metric	Analytical expression	Resilience metric	Analytical expression
Recovery time	$T_3 - T_0$	Performance loss	$\int_{T_0}^{T_3} \left(Q_0 - Q(t)\right) \mathrm{d}t$
Recovery rate	$\frac{\mathrm{d}Q(t)}{\mathrm{d}t} \mid T_2 \leq t \leq T_3$		
Deterioration rate	$\frac{\mathrm{d}Q(t)}{\mathrm{d}t} \mid T_0 \leq t \leq T_1$	ITAE	$\int_{T_0}^{T_3} t \cdot \lvert Q_0 - Q(t) \rvert \, \mathrm{d}t$
Initial impact	$Q_0 - Q_{min} \mid T_0 \leq t \leq T_1$		
Maximum impact	$Q_0 - Q_{min} \mid T_0 \leq t \leq T_3$	Profile length	$\int_{T_2}^{T_3} \sqrt{1 + \left(\frac{\mathrm{d}Q(t)}{\mathrm{d}t}\right)^2} \, \mathrm{d}t$
Residual functionality	$Q_{min} \mid T_0 \leq t \leq T_3$		
		Weighted sum	$\sum_{k=1}^{n} a_k \left(g(a_k) - Q(a_k)\right)$

2.6. Research gaps

The discussion of resilience definitions, evaluation approaches and quantitative measures in the previous sections led to the identification of the following research gaps with regard to the resilience of railway networks.

Research gap 1: The evolution of railway system performance during the consecutive resilience phases is not well understood for disruptions of varying scale and origin.

Under normal conditions, a railway network functions around a steady performance level, with only minor variations due to disturbances in the train service. When a disruption occurs, system performance can drop significantly. Performance during this period is represented by the resilience curve, also known as the bathtub model. Several articles were discussed in this chapter which quantified this behavior. All of those articles described a single or at most a few large-scale disruptions lasting for several days (e.g. Chan & Schofer, 2016) or months (e.g. Büchel et al., 2020), or disruptions that created an out-of-control situation (e.g. Janić, 2018). In reality though, disruptions of a smaller scale such as switch failures or signal failures occur frequently, while these have not been subject to resilience research. How system performance develops during these disruptions is therefore not well understood.

Research gap 2: Realization data have not been used to assess the resilience of a railway network for a large and heterogeneous set of disruptions.

Besides the lack of quantitative knowledge about railway system performance during disruptions, the simultaneous study and comparison of multiple disruptions constitutes a research gap from a methodological perspective. While modern technologies and data analytics create opportunities for the use of large amounts of empirical data in railways (Parkinson & Bamford, 2017), only Ren et al. (2020) collected data for a large and heterogeneous set of disruptions in the subway system. Wong et al. (2020) did so too for disruptions in air transport. Such extensive use of data to evaluate the performance during disruptions in railway networks has not been attempted before to the best of the author's knowledge, which is why it poses challenges with regard to data collection, preparation and analysis.

Research gap 3: The spatial attributes of a railway network have not been addressed explicitly when studying resilience as a function of time.

The spatial impact of disruptions has been addressed where appropriate in the storyline of this chapter. As the resilience curve is commonly defined as a function of time, the proposition was made to account for the spatial attributes of a system on the vertical axis of the graph, or alter-

natively, on a third axis. The articles that were reviewed have given insufficient insights into how the spatial attributes may actually be represented, because of a focus on the evolution of system performance in time rather than in time and space. Hence, the representation of the spatial attributes of a railway network forms a second methodological research gap.

2.7. Chapter summary

This chapter started with a general overview of railway resilience. The different system states and resilience phases were discussed and a revisited definition of resilience was given. Indicators for railway system performance were discussed and previous data-driven resilience evaluation approaches were explored for several transport modes. Also, metrics that describe the profile of the resilience curve were discussed. A theoretical research gap and two methodological research gaps were identified from the reviewed articles. The research gaps are addressed in the remainder of this report.

Answer to subquestion 1

With the knowledge obtained in this chapter, the first subquestion is answered.

> **Subquestion 1:** What can be learned from previous quantitative, data-driven approaches for resilience evaluation of railway networks?

Resilience research exhibits domain-independent characteristics, and therefore, resilience definitions and metrics are formulated in broad terms and could apply to general systems. This explains why the resilience metrics found in railway related articles also appear in general systems literature. Some additional metrics describing the nonlinearity of the resilience curve in the recovery phase are found in supply chain literature. These could be adopted for the resilience evaluation in railways as well. Performance indicators on the other hand are more domain-specific. With regard to performance measurement, it appears customary to compare the actual performance with planned performance, and also to aggregate the data, for example by day or geographical area. When the spatial propagation of disruption effects is studied, the spatial impact is commonly visualized in geospatial displays rather than in the resilience curve itself. In case resilience metrics or other quantitative measures are the main output of a study, statistical analyses are commonly performed to assess their significance or dependence, similar to the approach that is followed in this thesis. Regarding the data-driven resilience evaluation approaches in the reviewed articles, theoretical and methodological gaps exist by failing to study common types of disruptions, not realizing the potential of big data to assess network resilience and not explicitly addressing the spatial attributes of a system.

3. Practical background

The previous chapter described from a scientific viewpoint how disruption impacts and recovery may be evaluated. This chapter provides a more practical background on railway disruption management from the perspective of the infrastructure manager. This starts with a general overview. As the chapter progresses, more references are made to disruption management in the Netherlands. First, Section 3.1 describes in broad terms how disruption management is commonly organized in western Europe. Section 3.2 addresses the different types of disruptions that may occur and how they may be classified. Section 3.3 goes more in depth on how disruptions are managed in the Dutch railway network. Section 3.4 summarizes the chapter and provides the answer to subquestion two.

3.1. Disruption management in general

Railway traffic control is a complex task which is normally carried out by the infrastructure manager (IM). The workload of traffic controllers and dispatchers and the tasks they perform greatly depend on whether trains are running according to schedule or not. Traffic control in normal conditions (*besturing*) involves executing the traffic plan and process plan with the available planning, signaling and control systems. The traffic plan specifies the train paths, and the process plan specifies the operational utilization of infrastructure. However, as stated in Chapter 1, a disruption in the network requires additional input in the form of adjustments to the timetable, rolling stock and crew planning. Rolling stock and crew are rescheduled by the train operating company (TOC). Adjusting the timetable on the other hand is the responsibility of the IM. Traffic control during disruptions is therefore known as rescheduling (*bijsturing*). Disruption management is used as an umbrella term for the joint actions taken by the IM, TOCs and maintenance contractors during the rescheduling process.

Roles and task descriptions

Regardless of the country-specific organization of railway traffic control, a number of key roles are identified that are essential in the disruption management process. Commonly a distinction is made between regional and national control and between operations and traffic control by the TOCs and IM, respectively. The definition of "regional" control depends on the network structure and varies between countries or even within countries. Table 3.1 presents an overview of the key roles in traffic control based on Schipper and Gerrits (2018). Each row explains the role, the actor responsible for this role and the corresponding task description.

Table 3.1. Key roles and task descriptions in railway traffic control.

Actor	Role	Task description
IM	Train dispatcher	Allocate infrastructure capacity and monitor the safe movement of trains at a local level.
IM	Regional traffic controller	Monitor and optimize traffic flow at a regional level, assess decisions made by dispatchers and advise the national traffic controller.
IM	National traffic controller	Monitor and optimize traffic flow at a national level, support the regional traffic controller and advise the duty officer.
TOC	Regional operations controller	Monitor traffic flow, rolling stock and crew planning at a regional level.
TOC	National operations controller	Monitor traffic flow, rolling stock and crew planning at a national level.
TOC, IM	Team leader, shift leader, duty officer	Monitor the workload of controllers and maintain communication with other control centers.

Management actions

In the first phase of a disruption, emergency measures need to be taken and the affected area needs to be cleared in order to prevent an escalation of the situation. This is arranged by the traffic controller in cooperation with the dispatcher in charge of the affected area. A choice of measures, referred to as management actions, are available to controllers and dispatchers. Three main actions are identified: drive, reroute and cancel. Supplementary actions can also be taken. The available management actions are presented in Table 3.2. Each row specifies the action, its translation in Dutch and the description of this action. The actions that can be performed on a particular train depend on the train characteristics, the type of service (e.g. regional, intercity, freight) and the layout and location of a station in the network.

Table 3.2. Railway disruption management actions.

Action		Description
Drive	*Rijden*	The train is allowed to continue driving along its planned route.
Reroute	*Omleiden*	The train is diverted to arrive at its destination along a different route. Uncommon for regional and intercity services.
Cancel	*Opheffen*	The train is canceled and cannot start or continue on its route.
Short-turn	*Keren*	The train is rescheduled to head back in the opposite direction under a new train number. This usually does not require shunting.
Change track	*Spoor wijzigen*	The arrival track of a train is changed because it would otherwise result in a path conflict between two trains.
Retime	*Verleggen*	The departure time of a train is changed because it would otherwise result in a path conflict between two trains.
Reorder	*Volgorde wisselen*	The departure sequence of two trains is changed.
Add stop	*Stop toevoegen*	The train is rescheduled to make an extra stop.
Skip stop	*Stop overslaan*	The train is rescheduled to skip a planned stop.
Insert	*Inleggen*	The train is reinserted in the timetable when it has previously been canceled.

Tradeoffs

The described actions need to be coordinated between the IM and TOCs and across different levels of the organization, such as from planning to operation. How this is organized is defined by the underlying coordination structure. Schipper and Gerrits (2018) described a tradeoff between centralized and decentralized coordination. Centralized coordination involves operational decision making at a single location, whereas decentralized coordination involves decision making at dispersed locations. A decentralized structure makes the best use of local knowledge of the network and creates flexibility through direct communication and control over resources. However, a local optimal solution is not necessarily optimal for the larger network. A centralized structure would be preferred to keep a better overview of system performance, but it requires substantial information sharing. Given this tradeoff, a balance between centralized and decentralized coordination should be found.

A second tradeoff described by Schipper and Gerrits (2018) is the tradeoff between anticipation and resilience. With “resilience”, Schipper and Gerrits referred to the reactive capacity of controllers and dispatchers, not to be confused with resilience as defined in this report. Relying solely on the controllers’ expertise and reactive capacity may result in an ineffective response due to the dynamic and demanding situations controllers are faced with. For example, consider a disruption at a major station requiring all incoming traffic to be rescheduled. Alternatively,

the design of predefined solutions is characteristic of the anticipation approach, although not every disruption can be anticipated in detail. A minimum level of reactive capacity will always be required. Hence, a balance should be found for this tradeoff as well.

With regard to the above tradeoffs, considerable differences exist in the organization of railway disruption management in different countries. By understanding these differences, the relevance of the results from the data analysis can be determined for other countries. Schipper and Gerrits (2018) compared disruption management practices for the Netherlands, Germany, Austria, Belgium and Denmark. The performance of these countries on the discussed tradeoffs is illustrated in Figure 3.1. Regarding the first tradeoff, all countries have a division of tasks between national and regional control centers. However, the decentralized management by TOCs is unique to the Netherlands and Germany. In Germany this is due to the sheer number of TOCs. The size of the German network also means that decision making during disruptions is decentralized and that regional control centers have a greater level of autonomy compared to the other countries. Still, in all of these countries the IM and the dominant TOC are co-located in a national control center. In Belgium, the IM and TOC also work in mixed teams.

Regarding the second tradeoff, Schipper and Gerrits (2018) indicate that the reliance on predefined contingency plans in the Netherlands is high compared to the other countries. Austria has also developed plans for the most common disruptions, but these serve more as a template to controllers. The use of contingency plans is much less common in the other countries, which rely more on the reactive capacity of controllers to find a tailor-made solution to each disruption. In Belgium this is facilitated by the mixed and flexible composition of the teams, where other countries have special crisis rooms for shared decision making. In Germany, much communication also takes place over the phone, as face-to-face meetings with all the involved TOCs are difficult to arrange. All things considered, the Dutch and Austrian approach are mainly anticipatory, whereas decision making in Belgium is highly reactive. Whether this also results in a more resilient system remains an open question.

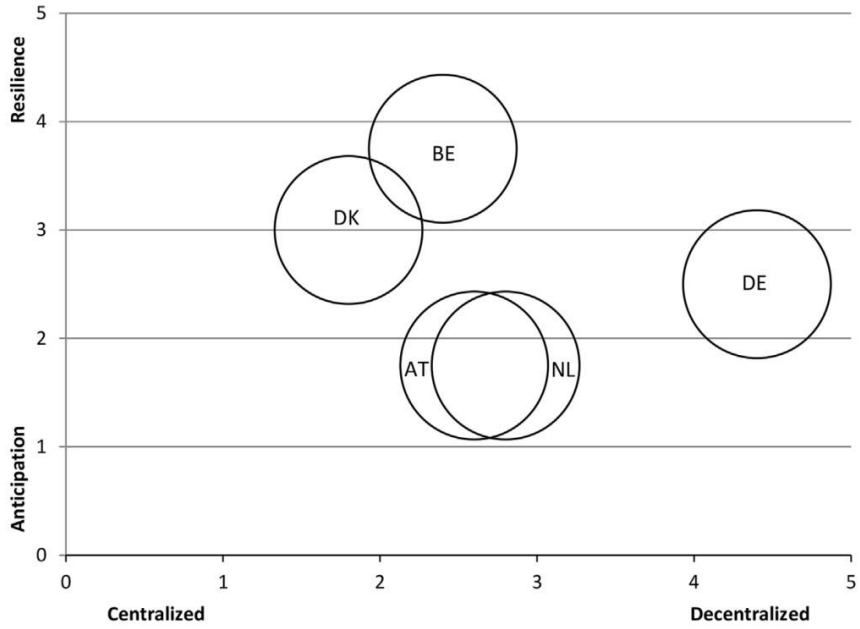


Figure 3.1. Disruption management tradeoff scores per country (Schipper & Gerrits, 2018).

3.2. Classification of disruptions

The response that is required by controllers and dispatchers depends on the type of disruption. It is therefore useful to understand which types of disruptions could occur and how they may be classified. Two types of classifications are discussed: by cause and by consequence.

Classification by cause

Although disruptions are inherent to transport systems, they are not random events. There is always a cause that triggers the disruption. Three subclassifications by cause are identified: by aggregated cause, specific cause and train incident scenario (TIS). First, classification by aggregated cause means tracing the cause of a disruption back to the rolling stock, infrastructure or an external event. Second, the aggregated cause can be specified further, which leads to a classification by specific cause. There are many possibilities for the specific cause of a disruption. Rolling stock related disruptions are mostly train defects, while infrastructure related disruptions are more diverse. These include the failure of signals, switches, overhead lines, crossings, bridges, etc. Even more diverse are disruptions with an external cause, as they include all remaining causes that cannot be influenced by the IM or TOC. Figure 3.2 presents the distribution of disruptions in the Dutch railway network (2,152 in total) for timetable year[1] 2019 per aggregated cause and per specific cause. The distribution per aggregated cause is fairly even, but the distribution per specific cause shows a greater differentiation. The top five specific causes occurring at least once per week on average are train defects, section or signal failures, collisions[2], switch failures and overhead line failures. All remaining specific causes were labeled as "other causes".

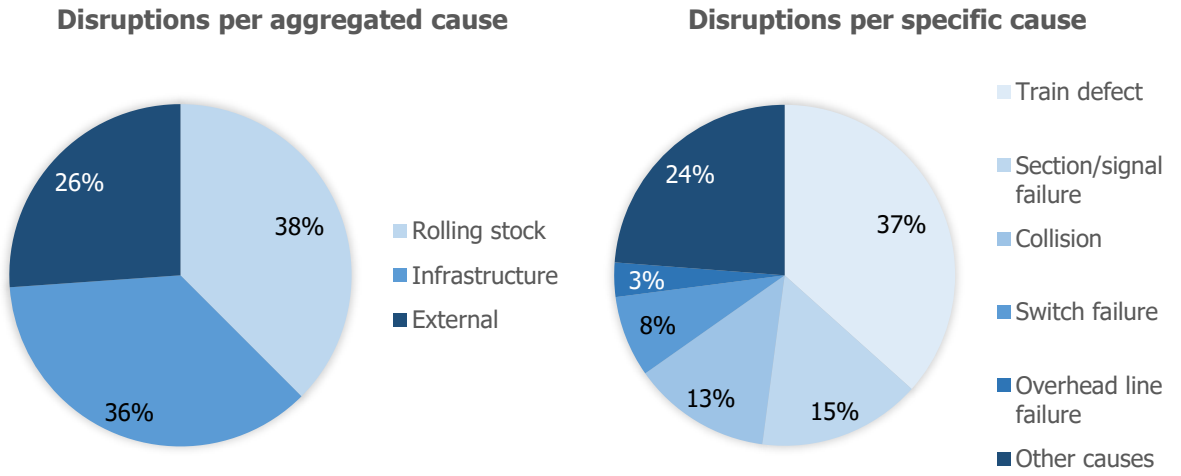


Figure 3.2. Distribution of disruptions per aggregated cause (left) and specific cause (right) for 2019.

Third, disruptions may be classified by cause based on the train incident scenario registered by ProRail. This classification is specific to the Netherlands. The TIS prescribes the response and eventual upscaling by the IM, TOCs and emergency services, which may vary per region or location. Similar scenarios exist for incidents occurring on highways, on waterways and in air transport. From the perspective of traffic control, a TIS is purely administrative*. Five TIS categories have been specified:

- TIS 1: Disrupted train service
- TIS 2: Fire
- TIS 3: Collision or derailment
- TIS 4: Hazardous goods
- TIS 5: Bomb threat

[1] A timetable year starts the second Sunday of December and ends the second Saturday of December the year after. Timetable year 2019 lasted from December 9, 2018 until December 14, 2019.
[2] Includes collisions with a person, (motor) cyclist, small road vehicle, large road vehicle, large animal and infrastructure object.

Each category has four subcategories depending on the severity of the event. Many disruptions (85% of the cases in 2019) are classified as TIS 1, although their causes may differ significantly. As a result, classifying disruptions by TIS would create a diffuse picture of system performance during disruptions which is not preferred for the data analysis.

Classification by consequence

Apart from classification by cause, disruptions may be classified by their consequence. Two subclassifications are identified: by impact type and by customer hindrance. Some new definitions are introduced first:

- Timetable point: A location in the railway network for which a plan time is included in the timetable. Usually a train station, but also possibly a bridge, junction, shunting yard entrance, etc.
- Line: Railway infrastructure between two timetable points consisting of one or multiple sections.
- Logistical functionality: The ability to accommodate stranded passengers, accommodate crew, park stranded or malfunctioning trains, or at least allow trains to short-turn.
- Boundary point: A timetable point that can be marked as the boundary of a disrupted area. Amounts to the closest timetable point with logistical functionality relative to the location of a disruption. A map of all boundary points is included in Appendix E.
- Decoupling point: A timetable point where trains are allowed to start or end their route in case of a disruption.

First, disruptions are classified by consequence based on the impact type, which determines the remaining infrastructure capacity. Three impact types are identified: line blockage (partial or full), timetable point outage (partial or full) and control center outage. In addition, a line or timetable point may experience reduced functionality, for example because of temporary speed limitations or the deployment of emergency services. Classification by impact type is a common approach in the Netherlands. A description of the impact types is presented in Table 3.3. Each row specifies the impact type, its translation in Dutch and the description of the impact type. Figure 3.3 presents the distribution of disruptions per impact type, which shows that 85% of the disruptions in 2019 were line blockages and a mere 10% were timetable point outages. Control center outages are so uncommon that they do not appear in the chart.

Table 3.3. Disruption impact types based on infrastructure availability.

Impact type		Description
Reduced line functionality	*Functiebeperking baanvak*	All tracks on a line are available, but the capacity is reduced.
Reduced timetable point functionality	*Functiebeperking begrenzingspunt*	All tracks at a timetable point are available, but the capacity is reduced.
Partial line blockage	*Partiële baanvakstremming*	One or multiple tracks on a line are blocked. At least one track remains available.
Full line blockage	*Volledige baanvakstremming*	All tracks on a line are blocked. There is no train traffic on the particular line.
Partial timetable point outage	*Partiële begrenzingspuntuitval*	One or multiple tracks at a timetable point are blocked. At least one track remains operational.
Full timetable point outage	*Volledige begrenzingspuntuitval*	All tracks at a timetable point are blocked. There is no train traffic to or from the particular timetable point.
Control center outage	*Postuitval*	The functionality of a regional control center is compromised. There is no train traffic in the entire traffic control area.

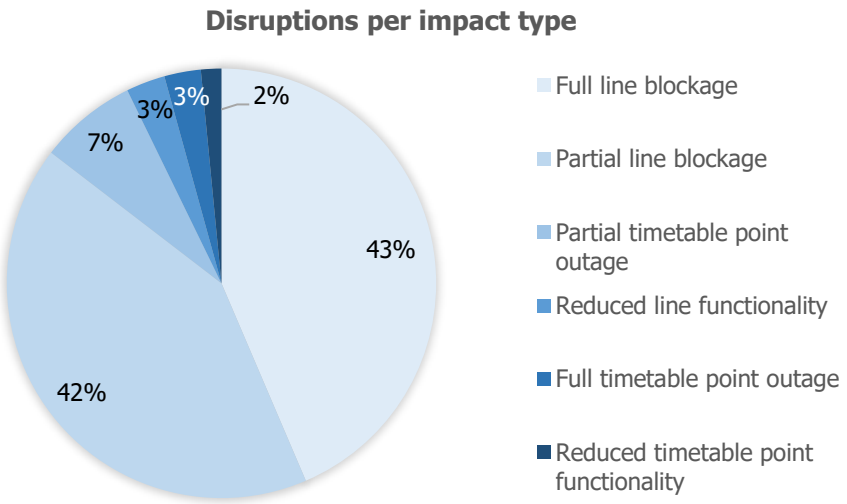


Figure 3.3. Distribution of disruptions per impact type for 2019.

Second and more specific to the Netherlands, disruptions are classified based on customer hindrance (*klanthinder*), which amounts to the cumulative delay in minutes caused by a single disruption. The train passenger is defined herein as the customer. Each disruption is allocated to one of four hindrance classes (*hinderklasse, HK*):

- HK 1: Total delay ≥ 2400 min.
- HK 2: 680 min. ≤ Total delay < 2400 min.
- HK 3: 40 min. ≤ Total delay < 680 min.
- HK 4: Total delay < 40 min.

Customer hindrance is calculated after the end of a disruption to evaluate the impact of the disruption on the train service. For trains with a delay greater than 5 minutes, the actual delay is used in the calculation. For rerouted trains, a standard 15 minutes are added to the cumulative delay. For canceled trains, a standard 30 minutes are added. A maximum yearly number of HK 1 and 2 disruptions is agreed upon between the Dutch government and ProRail, which is why customer hindrance should be prevented as much as possible. If the agreed number is exceeded, a fine is imposed by the government. The way in which customer hindrance is included in the design of contingency plans and during rescheduling is explained in Section 3.3.

Based on the discussed classifications, it is expected that classifying disruptions by their specific cause yields the most uniform picture of system performance within each group, as classification by aggregated cause or TIS is too generic and classification by consequence would create a rather unequal comparison. After all, line blockages are far more common than timetable point outages or control center outages. The same holds for HK 3 disruptions, which are far more common than HK 1 or 2.

3.3. Disruption management in the Netherlands

The way in which disruption management is organized in the Netherlands is specific to the country and permeates through the data used in the data analysis. Therefore, the organization and working practices need to be properly understood in order to correctly interpret the data, draw valid conclusions from the experiments and make useful recommendations based on the results. This section discusses the coordination structure, process flow, design and application of measures, identification of impact areas, performance measurement and state of the practice regarding disruption management in the Netherlands.

The Dutch railway traffic coordination structure

As mentioned in Section 3.1, railway traffic control and operations control in the Netherlands are both characterized by a division of tasks between national and regional control centers. Traffic is coordinated nationally from the Operational Control Center Rail (OCCR) in Utrecht, which facilitates cooperation between the IM, TOCs and maintenance contractors. An overview of the coordination structure is shown in Figure 3.4. Traffic control at the OCCR is the responsibility of the Central Monitoring and Operations Control Center (*Centraal Monitoring en Beslisorgaan, CMBO*) which employs four national traffic controllers (*verkeersleider CMBO, VLC*): three for passenger traffic and one for freight traffic. The VLCs are supervised by the duty officer rail (*officier van dienst spoor, OvD-S*). Requests for unscheduled maintenance are managed by a planner. When necessary, the duty officer may contact the supervisor of NS' national control center, which otherwise plays a minor role in disruption management.

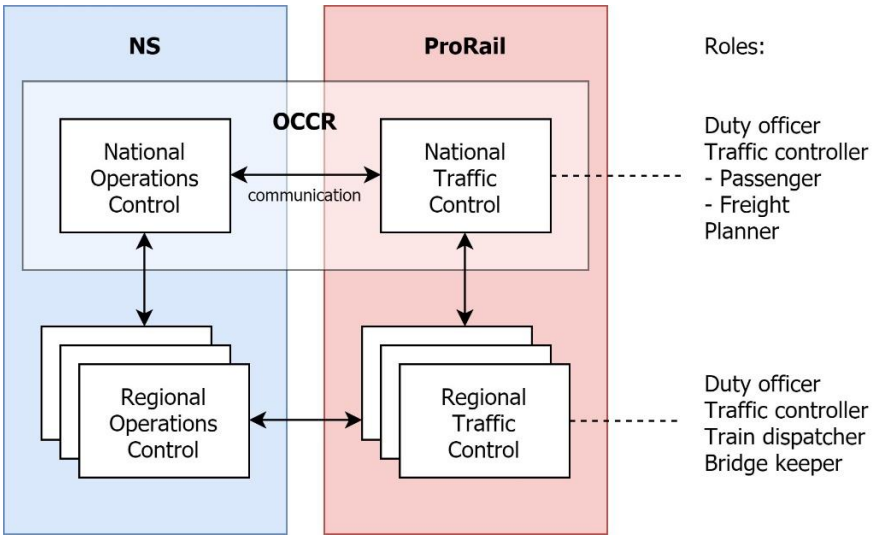


Figure 3.4. The Dutch railway traffic coordination structure.

Regional traffic control is distributed over twelve regional traffic control centers, each of which are operated by one or two traffic controllers (*decentrale verkeersleider, DVL*) and multiple train dispatchers (*treindienstleider, TRDL)* supervised by a duty officer. Some control centers also have a bridge keeper employed. Depending on the location in the network, regional traffic control handles a variety of TOCs. Control centers in the middle of the country are mostly faced with NS traffic, while control centers in the periphery are more involved with regional TOCs. Control center Kijfhoek is an exception as it exclusively deals with freight traffic. A map of the areas monitored by the control centers is included in Appendix F. The dominant TOC in the Netherlands, the NS, also has a decentralized coordination structure. Operations control is distributed over five regional operations control centers (*regionaal besturingscentrum, RBC*). Communication between ProRail's regional control centers and NS' RBCs is maintained for the purpose of rolling stock and crew rescheduling.

Process flow and decision making

The process that is initiated when a disruption occurs is large and complex, because it involves many process steps and decisions by multiple actors. During this process the bathtub model is followed. The model is fundamental to the operation of traffic control in the Netherlands, and therefore, each bathtub phase (or: resilience phase) is characterized by a unique set of tasks. A simplified overview of the process flow from the perspective of traffic control is shown in a swimlane diagram in Figure 3.5. The sequence of events is portrayed vertically and starts at the top of the diagram. The diagram offers an updated and more complete overview compared to Ghaemi et al. (2017). For the sake of readability, decision boxes were not included.

First phase
The process starts with a notification, usually reported by a train driver or an external source. The notification is registered by the back office of the control room (*meldkamer spoor, MKS*) at the OCCR. The control room distributes the notification and appoints a general controller (*algemeen leider, AL*), who may be sent to location to assess the situation. When necessary, mechanics are sent to location. After gaining a first perception of the situation, initial measures are taken by the dispatcher and the DVL regarding safety and logistics, respectively. The DVL is supported by the VLC who may take additional measures to ensure that timetable points outside the disrupted area do not become saturated. After gaining a detailed perception of the situation by the dispatcher and the DVL, the DVL proposes a capacity reallocation (*verdeelbesluit, VDB*) which is checked by the VLC. A capacity reallocation specifies the boundary points and impact type of the disruption, and with that, also the remaining infrastructure capacity. Based on the capacity reallocation, a contingency plan (*versperringsmaatregel, VSM*) is selected by the DVL which is again checked and possibly modified by the VLC. A VSM prescribes the actions to be taken for the relevant train series in the disrupted area as specified in Table 3.2, and it essentially provides a revisited traffic plan for the second phase. In case no suitable VSM is available, the next most suitable one is selected and modified. Before the VSM is applied, the disrupted area must be cleared by the dispatcher and the DVL. This typically involves retrieving stranded trains. The TOC is informed as the preparation for the VSM may require trains to be shunted.

During the first phase, the workload on dispatchers and controllers is highest. This phase is characterized by much communication and the time pressure to prevent an out-of-control situation. A complicating factor is that ProRail has little to no insight into crew rescheduling by the RBC*. Also, the first phase becomes complicated when incorrect or fragmented information is received by traffic control. For instance, it could be the case that a stranded train is running again while measures are already being taken, which then causes substantial delays*. A characteristic of the first phase that could be observed in the resilience curve is the “drain” of the bathtub, which represents the lowest point in the curve. This may occur when trains are halted as a result of emergency measures. Performance is then slightly restored towards the end of the first phase, before the VSM is applied*. Another characteristic of the first phase is wishful thinking: the tendency to believe that a disruption will resolve itself. This may cause trains to keep running with minor delays, potentially creating a disruption elsewhere in the network due to the accumulation of delays*.

Second phase
The start of the second phase represents the transition from disrupted traffic to traffic that follows the revisited timetable, eventually leading to a steady state. In practice, this transition will be smooth rather than immediate. The second phase starts with monitoring the revisited traffic plan by the dispatcher, DVL and VLC. Meanwhile, the VLC monitors if the disruption is at risk of becoming a HK 2 disruption. If that is the case, the OvD-S is informed. The OvD-S may arrange a meeting to discuss additional measures. Critical at this stage is to check whether the capacity reallocation is as well defined as possible. The DVL logs mutations to the traffic plan for the first 30 minutes of the VSM in the computer system. Subsequent mutations are logged by the VLC. Apart from monitoring traffic in the current state, the second phase involves preparing a plan to restart the train service according to the original timetable. First, remaining obstacles such as delayed freight trains must be cleared. The sequence in which train series are reinserted is usually included in the VSM. If not, the VLC checks this with the involved TOCs. The DVL further prepares the restart plan in detail and checks with the dispatcher and the TOC if the plan is feasible. When the plan is approved and the infrastructure is reclaimed by the AL, the VLC announces that the restart can be initiated. The definitive prognosis for this moment should be made 30 minutes in advance by the AL. In practice, it may occur that the restart is

already possible before prognosis, in which case the VLC is taken by surprise because a feasible restart plan has not yet been devised*.

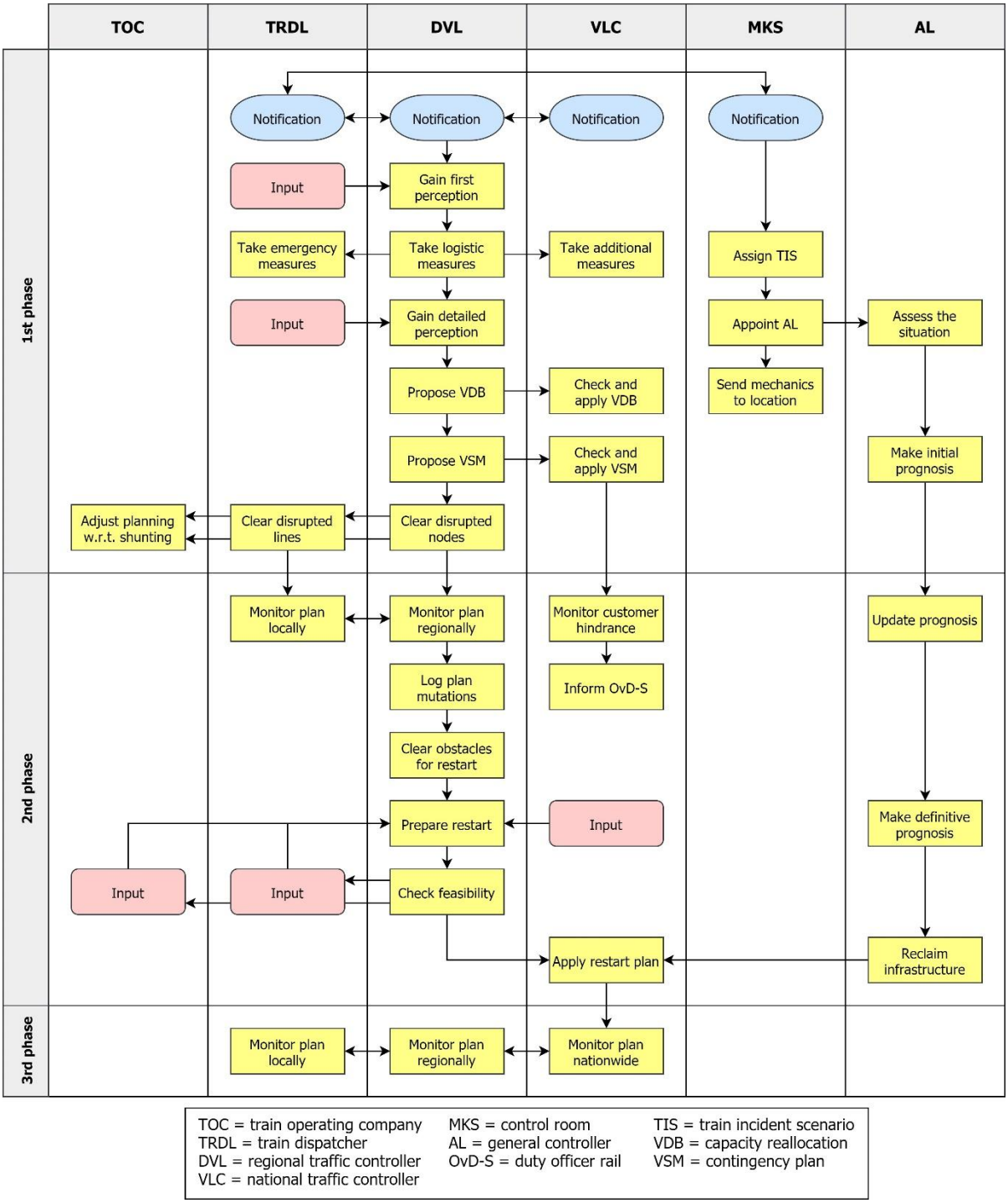


Figure 3.5. Swimlane diagram of the main disruption management processes.

Third phase

In the third phase, the restart plan is applied and monitored. The purpose of the restart plan is to return to the original timetable. At this point, ProRail's control over the situation is reduced compared to the previous two phases, since there is not much more to be done from the IM's perspective. The infrastructure has been made available, and it is up to the TOCs to get the trains running again. Naturally, the restart is postponed if the TOCs are not ready. Similar to the first phase, NS' crew planning can be a bottleneck when initiating the restart,

because crew is taken off the train when the train is canceled and put back on the train prior to restart. This is not the case for regional and freight TOCs, which typically prefer to let their crew stay with the delayed or canceled train due to less flexibility in rescheduling*.

Design and application of measures

The process flow described in Figure 3.5 reveals that the application of predefined solutions is crucial to the disruption management process. In literature (e.g. Veelenturf, 2014; Ghaemi et al., 2018), these solutions are referred to as contingency plans. For the design of such plans, the network has been divided on corridor level. Every line in the Dutch railway network (except for the *Betuweroute*, which is freight only) is part of one of four passenger corridors, which are the *A2/A12*, *IJssellijn*, *Oudelijn* and *Groene Hart/Veluwe*. A map of the corridors is included in Appendix G. The division on corridor level lays the foundation for the design of measures for train delay management (TADs) and disruption management (VSMs), which are discussed separately in this subsection.

Train delay management

Traffic control on corridor level requires a shared perspective between controllers in different control centers on how to handle delays. This is documented in a guideline (*leidraad*) for train delay management unique to each corridor. A guideline is a document of a few pages describing the goals, principles and measures on corridor level. The first guideline was developed jointly with the NS for the A2/A12 corridor in anticipation of the frequency increase between Amsterdam and Eindhoven. Guidelines for the other corridors followed shortly after the first one based on the corridor-specific goals. For example, the goal on the A2/A12 corridor is to maintain the 10-minute rhythm, whereas the goal on the IJssellijn is to facilitate as many connections as possible to the intercity service that traverses the corridor. The measures that are described in the guideline vary accordingly. There are currently no guidelines for regional and freight routes because of the lower complexity and demand*.

Specific measures and dispatching rules for handling delays are defined in train delay handling documents (*treinafhandelingsdocument, TAD*) unique to each dispatch area. A TAD describes how long a train is allowed to wait for a connecting train and how a train series needs to be rescheduled or canceled when delays grow too large. For example, the TAD for dispatch area Eindhoven specifies that the 3900 series from Heerlen to Amsterdam shall be canceled when it has a delay of 10 minutes or more by the time it reaches Roermond. The TADs are built on existing knowledge from before the guidelines and current VSMs were developed*. Nowadays, the guidelines provide a framework for updating and improving TADs, and they prescribe a coherent approach in case a TAD does not offer resolution.

Disruption management

Because TADs are not suitable for use in disruption management, VSMs are designed by the VGB team. The philosophy behind the VSMs is documented in an assessment framework that outlines the interests and prioritization of different service types. The framework provides rules for scenario makers who design and manage the VSMs and TADs, and more simple rules for traffic controllers to be followed when no VSM is available. Essential in both cases is the allocation of remaining infrastructure capacity when capacity is reduced due to a disruption. The standpoint that ProRail takes in this respect is twofold. Of the two principles, the first one is superior to the second one:

1. Infrastructure should be utilized to maximum capacity.
2. TOCs should be treated nondiscriminatory.

The starting points in VSM design are that measures should be predefined based on the TOCs' interests and that they should be feasible. This means a VSM has to meet planning standards (i.e. headway and crossover times) as well as minimal process times to ensure the safe opera-

tion of trains. Planning standards are generic, although an additional calculation can be made which takes location-specific characteristics of the infrastructure or rolling stock into account. This may result in a tighter, but still feasible VSM that utilizes the potential of the local infrastructure as much as possible. However, the tight planning in a VSM is also a potential source of new delays*, which could be observed in the response phase of the resilience curve.

Because a VSM is designed as a revisited timetable, it follows a basic hour pattern, which is a generic timetable for exactly one hour that can be repeated throughout the day. The translation from basic hour pattern to an actual timetable and from train series to train number needs to be made in real time by traffic control. Based on the number of reroutes and cancellations in the VSM, customer hindrance can be calculated in advance. The calculated cumulative delay is increased by 10% to account for delays occurring during a disruption, which cannot be anticipated with certainty. Still, customer hindrance has little added value in VSM design because the scenario maker builds security into the VSM* to ensure that the VSM will nearly always be feasible. Relaxation of the VSM can be considered in real time by the duty officer to prevent customer hindrance of HK 2 or worse.

Each VSM is designed to accommodate a specific impact type as presented in Table 3.3. This means for example that a line blockage is treated equally, regardless of whether it is caused by a collision or an overhead line failure. However, the effects and duration of these events could differ significantly. VSMs are mostly designed for line sections, since line blockages are more common and easier to deal with than disruptions at timetable points, especially when a point serves more than two directions. This approach yields a decent coverage*, although the end goal remains to develop a suitable plan for every kind of infrastructure restriction. The current VSMs have only been designed and updated in recent years based on the principles from the Railway Control and Rescheduling of the Future program (*Be- en Bijsturing van de Toekomst, BBT*). A core principle of BBT is to prevent cascading effects, which means in practice that a single train is inferior to the larger network: it is preferred to cancel a delayed train rather than to keep the train running with a delay that could eventually propagate onto successive trains.

Identification of impact areas

The actions included in a VSM not only vary per train series, location and impact type, but also per impact area. ProRail makes a distinction between the first, second and third impact area. The first impact area is bounded by the first intercity (IC) decoupling points from the location of the disruption. The impact on the train service is greatest in this area. The second impact area is bounded by the next closest IC decoupling points from the first decoupling points. This means that the complexity of the impact area so far increases quickly for a strongly connected network, and the size of the impact area increases quickly for a location in the network with a low decoupling point density. The third impact area again is bounded by the next closest IC decoupling points from the second decoupling points. The impact on the train service should be minimal in the third impact area. This is underlined by the fact that, in principle, it is not allowed to cancel trains in the third impact area, whereas cancellations are allowed in the first and second impact area as long as certain conditions are satisfied.

Figure 3.6 illustrates how the impact area is determined per impact type for part of a hypothetical network. A real-world example is provided in Chapter 5. The figure provided here assumes a worst-case scenario since according to the newest definition, a decoupling point only becomes part of an impact area when it is reached by a “contaminated train”, which is a train that passes through the disrupted line or timetable point. In Figure 3.6, the orange and yellow colored points would not be part of the impact area if all disrupted traffic were to be contained in the first impact area. Hence, in the newest definition, the impact area is not geographically fixed based on the location of a disruption, but it is codependent on the remaining traffic flow. In this thesis, the old definition (described in the previous paragraph) is followed.

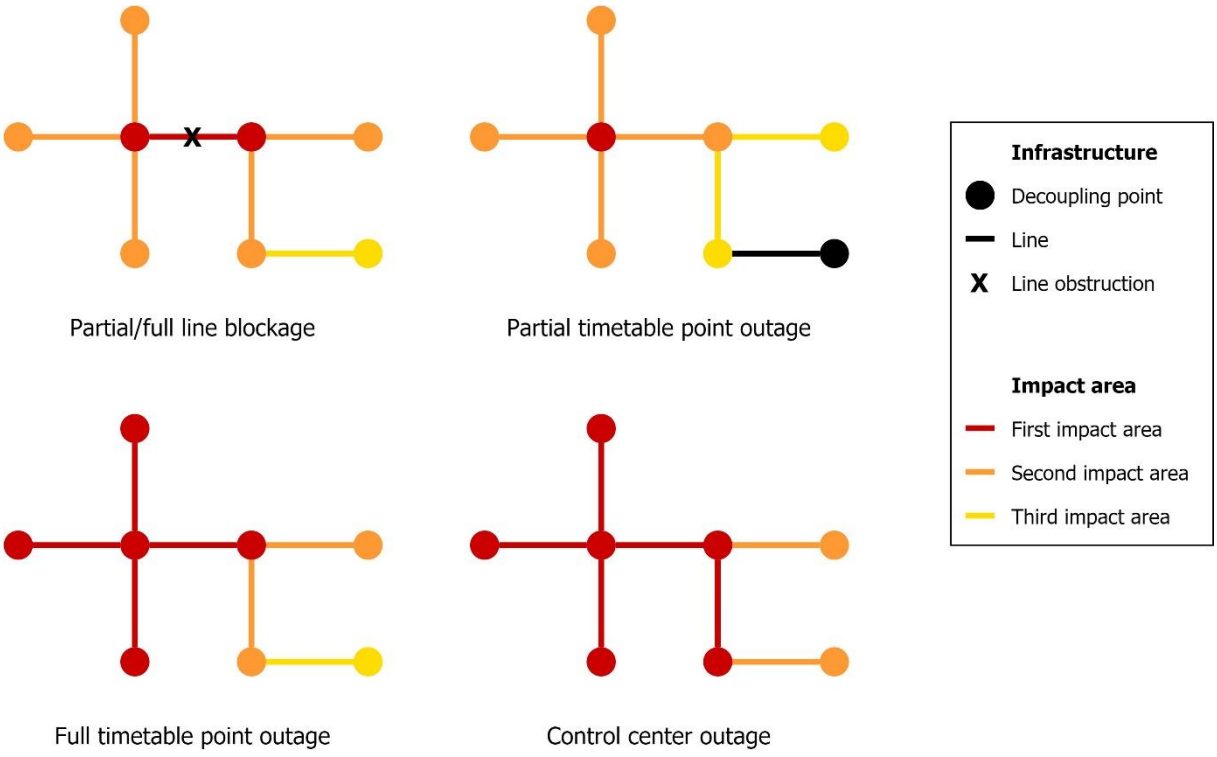


Figure 3.6. First, second and third impact area per impact type.

Performance measurement

Since the bathtub model plays such a prominent role in disruption management practices in the Netherlands, one might expect that close attention is paid to performance measurement along both axes of the curve, but this is not the case. Rather, the model serves mostly as a theoretical foundation. Traffic control during disruptions is assessed on six critical performance indicators (KPIs) which all describe lead times in the first phase. There are no KPIs for traffic control in the second and third phase. The KPIs apply to national traffic control by the CMBO, but not to traffic control in the regional control centers. Furthermore, they relate to the internal processes, not to system performance as shown in the resilience curve. The KPIs for the CMBO describe the following time differences:

1. "Notification known to MKS" until "concept VDB and VSM"
2. "Concept VDB and VSM" until "definitive VSM"
3. "Definitive VSM" until "VSM applied"
4. "VSM applied" until "cancellations until expected end of disruption logged by VLC"
5. "VSM applied" until "framework for restart logged by VLC"
6. "VSM applied" until "traffic according to VSM"

When all KPIs are met, the first phase should take no more than 44 minutes, which equals the time until a notification reaches the control room plus the summed target values for KPIs 1, 2, 3 and 6. The KPIs are monitored, but they are not a hard constraint. Additional performance indicators are available that do relate to railway system performance, but those indicators are not considered critical. An overview of these indicators and their definitions is given in Table 3.4. One of the indicators in the table (passenger traffic punctuality) has been a KPI for ProRail in the past, but was dropped after passenger punctuality had been redefined by the NS in 2017. Note that passenger *traffic* punctuality indicates the punctuality of trains, whereas passenger punctuality (based on check-in times) indicates the punctuality of people's travels. Passenger traffic punctuality is still reported and used to assess the performance of regional services*.

Table 3.4. System performance indicators used by ProRail.

Performance indicator	Definition
Realized train paths	The percentage of realized train paths for all TOCs. A train path is realized when an activity is registered at all timetable points of this path, either by the original train or a replacement train.
Delivered train paths	The percentage of realized train paths for all TOCs plus all unrealized train paths that are caused by the TOCs.
Passenger traffic punctuality	The percentage of passenger train arrivals with a delay of less than three minutes.
Freight traffic punctuality	The percentage of freight trains that reach their final destination with an additional delay of less than three minutes relative to the delay at the starting point.
Canceled services	The percentage of train activities included in the original traffic plan that have not been realized.

Along with the change of passenger traffic punctuality to a noncritical performance indicator, the so-called "red" and "black" days lost their official status as well. Yet, it remains common practice to refer to a traffic day in the following manner:

- Green day: A traffic day characterized by a traffic punctuality greater than 92.5% and canceled services lower than 1%.
- Red day: A traffic day characterized by a traffic punctuality lower than 85% and/or canceled services greater than 5%.
- Black day: A traffic day characterized by a traffic punctuality lower than 75% and/or canceled services greater than 10%.

Red and black days are often related to extreme weather and can thus be anticipated based on the weather forecast. A red day may also be the result of an unfortunate or incorrect decision by traffic control or by problems with the TOC (e.g. a shortage of shunting train drivers). Sometimes, a red day may simply be coincidence, when multiple disruptions lead to the congestion of a major node in the network*. Red and black days present a challenge to traffic controllers, for on those days, disruptions can no longer be considered independent from each other. This means VSMs need to be adjusted to meet the current state of traffic in the network, which takes more time and communication than usual* causing a longer first phase. A timeline of red and black days and some of their causes for calendar year 2019 is presented in Figure 3.7. The black day of November 27 stands out as no single cause could be identified. This simply appears to have been a rather eventful day, with an overhead line failure on a single track line between Leiden and Utrecht and a train defect occurring shortly afterwards at the station in Woerden severely disrupting railway traffic in the western part of the country.

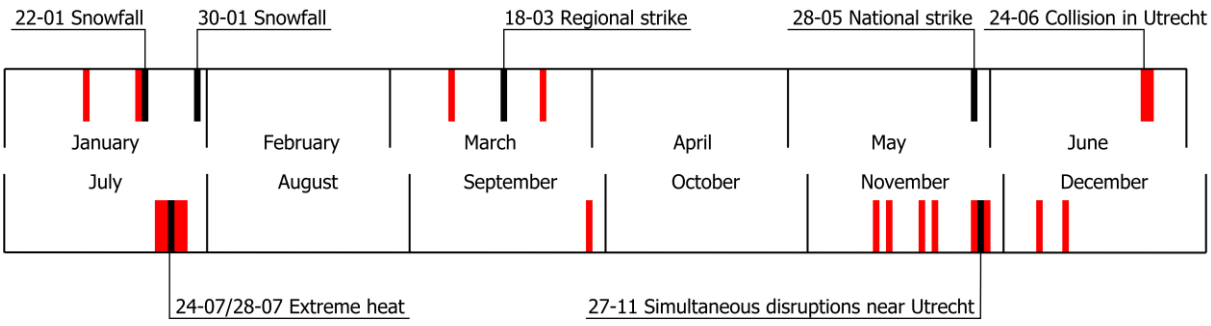


Figure 3.7. Timeline of red and black days in 2019.

State of the practice

Although the bathtub model and the resilience curve are two different names describing the same concept, the resilience of the railway network has not been studied by ProRail before. Improvement initiatives have been undertaken based on the bathtub, but those studies focused on a single phase of the curve with the goal to find process improvements. For example, a study on the first phase identified the elements that may contribute to a complex disruption with a long first phase and proposed ways to act on this. A study on the third phase found a positive correlation between the duration of the third phase and the likelihood of causing customer hindrance of HK 1 or 2. Furthermore, a dashboard on the bathtub phases has been developed in Microsoft Power BI. The dashboard reports the average duration of (parts of) each phase. The total duration is also reported for the most common types of disruptions. Other than that, existing knowledge about resilience is limited and provides little foundation to build on.

3.4. Chapter summary

This chapter started with an overview of disruption management practices regarding the roles and responsibilities, management actions and tradeoffs in disruption management. Possible ways to classify disruptions by cause and by consequence were discussed. Disruption management in the Netherlands was discussed with an emphasis on the process flow, design of predefined solutions, identification of impact areas and measurement of system performance.

Answer to subquestion 2

With the knowledge obtained in this chapter, the second subquestion is answered.

> **Subquestion 2:** What is the current state of the practice and quantitative knowledge regarding different types of railway disruptions in the Netherlands?

Each country has its own specific coordination structure and approach to railway disruption management. In the Netherlands, there is a heavy reliance on predefined solutions known as contingency plans. These plans specify the actions to be performed on each train series in the disrupted area and are designed to function as a revisited timetable. Contingency plans apply to the second phase of the bathtub model. The bathtub model is rooted in ProRail's working practices regarding disruption management, and each phase in the model is characterized by a unique series of processes. These processes require cooperation between the control room, dispatchers, traffic controllers, train operating companies and maintenance contractors.

Contingency plans consider the effects on infrastructure capacity, but not the underlying cause of a disruption. The aggregated causes are rolling stock, infrastructure and external, but these can be specified further. In the light of resilience, not much is known about different types of disruptions other than the average duration of each phase and the total duration of a disruption for the most common causes. Although these durations are known, they are currently not used to evaluate the resilience of the network. As a consequence, the evolution of system performance during disruptions is not known, even though the necessary traffic realization data are available. This means there is a similar knowledge gap in practice as there is in scientific literature regarding the quantitative knowledge about different types of disruptions, or even disruptions in general. This gap is addressed in the upcoming chapters.

4. Methodology

The previous two chapters addressed the theoretical and practical background on resilience and disruption management in railways. Building on the insights obtained in those chapters, this chapter describes the methodology for the data analysis. Section 4.1 discusses the performance indicators and resilience metrics that were selected in order to quantify railway system performance. Section 4.2 presents the resilience evaluation framework and discusses the input data and calculation procedure in detail. Section 4.3 discusses the algorithms that were developed to calculate the resilience curve and resilience metrics. Section 4.4 gives an overview of the statistical methods to be considered and the theory behind them. Section 4.5 summarizes the chapter and provides the answer to subquestion three.

4.1. Selection of indicators and metrics

A crucial first step prior to data analysis is the selection of one or more performance indicators by which to calculate the resilience curve. Also, a selection should be made of resilience metrics that describe the profile of the resilience curve quantitatively. The selection of indicators and metrics is covered in this section.

Performance indicators

In Chapter 2, four categories of real-time performance indicators were identified: travel time, travel demand, traffic flow and use of resources. Such categories are referred to by Jafino et al. (2020) as "functionalities". In Chapter 3, five additional performance indicators used in ProRail were presented that partly resemble the previously identified indicators. Unique additions were the *realized* and *delivered train paths*, but since a train path spans a period of time and can therefore only be evaluated after the fact, these indicators were not considered as viable alternatives to the available indicators. As a result, sixteen potential indicators were considered, accounting for the various definitions of punctuality. Empirical similarity between the indicators should be avoided, which means no two indicators of the same functionality should be selected (Jafino et al., 2020). An updated overview of the available performance indicators per functionality is presented in Table 4.1 based on Table 2.3.

Table 4.1. Available system performance indicators per functionality.

Travel time	Travel demand	Traffic flow	Use of resources
Passenger punctuality	Satisfied demand	Traffic throughput	Track occupation
Passenger traffic punctuality	Passenger volume	Speed reduction	Rolling stock usage
Freight traffic punctuality	Network saturation		Transport capacity
Journey time			Canceled services
Average delay			
Secondary delay			
Cumulative delay			

Some of the issues with certain indicators were already addressed in Chapter 2. First, all travel demand indicators require passenger counts or at least accurate estimates, which are typically not available to the infrastructure manager and particularly not in real time. Second, the speed reduction and rolling stock usage are difficult to retrieve as this kind of information lies with the TOCs and may be considered as competitively sensitive information*. In terms of traffic flow indicators, traffic throughput is therefore the better option compared to speed reduction, although it is more appropriate for studying a route or cross-section than an entire area. In terms of resource indicators, data access is less of an issue for track occupation and transport capacity compared to rolling stock usage. However, the problem with these indicators is that their direction is ambiguous: a higher value generally indicates better performance, but this

switches around when trains are halted*. Canceled services is a more suitable resource indicator as its direction is unambiguous: a cancellation indicates nonperformance by definition.

In terms of travel time indicators, journey time is the least suitable given the lack of a consistent reference value. A slightly better option is the secondary delay, which is difficult to retrieve as it requires disentangling primary and secondary delays. The three best options are the average delay, cumulative delay and punctuality. Regardless of its exact definition, punctuality has the disadvantage that it is binary: a train activity is either punctual or not, based on a tolerance for the maximum acceptable delay. It does not tell whether the delay is for example 5 minutes or 20 minutes. However, an advantage of punctuality is that it is the most relatable indicator to visualize in a resilience curve, as it represents an increasing system service function with clear upper and lower bounds. For the average and cumulative delay to be described in a similar way, delay measurements would have to be transformed and normalized. While this is not an issue when studying an individual disruption, it becomes problematic when studying multiple disruptions simultaneously, as the normalized delay is not comparable across disruptions. This means that punctuality is ultimately the preferred travel time indicator.

To summarize, two performance indicators were preferred over the rest: canceled services and punctuality. These are complementary since punctuality does not include cancellations, and canceled services does not include delays. The two indicators shall therefore be used alongside each other*. To allow an easy comparison between them, both indicators should preferably represent an increasing system service function. Therefore, canceled services was transformed into traffic intensity, which represents the percentage of realized train activities. Hence, traffic intensity is effectively the opposite of canceled services. With respect to punctuality, the traffic punctuality was considered since passenger punctuality data are not available to ProRail. Also, traffic punctuality has the advantage that it is not affected by anticipating passengers and replacement transport, which is a problem with passenger punctuality*. For example, when a passenger completes the first leg of a trip by train, switches to a replacement bus and completes the last leg of the trip by train again, this could be interpreted as two punctual trips although the total travel time is longer than usual. In line with the definition of traffic punctuality in ProRail, a three minute tolerance was set for the maximum allowable delay. However, as a deviation from this definition, punctuality was defined for arrivals as well as passings. This was done to capture the real-time state of the railway system as accurately as possible. Ultimately, this led to the following definitions of railway system performance.

Traffic punctuality: The proportion of train activities with a delay of less than three minutes relative to the number of realized train activities.

Traffic intensity: The proportion of realized train activities relative to the total number of planned train activities in the timetable.

To include both indicators in the same curve, they were combined in a composite performance indicator Q which is the weighted sum of traffic punctuality (P/R) and traffic intensity (R/T):

$$Q = \left((1 - \lambda) \frac{P}{R} + \lambda \frac{R}{T} \right) \cdot 100\% \tag{2}$$

Where:

- P is the number of punctual train activities
- R is the number of realized train activities
- T is the number of planned train activities
- λ is the normalized performance weight ($0 \leq \lambda \leq 1$)

It is acknowledged that there are downsides to using a composite indicator. First, the result is somewhat abstract and less easy to communicate than when using a single indicator. Second, there is an implicit relationship between the two indicators that cannot be ignored, since P is a subset of R and R is a subset of T. The dependency between the indicators is reduced since the proportions are evaluated, but the implicit relationship remains. Yet, there are also benefits to using a composite indicator. Most importantly, the composite indicator is useful to represent delays and cancellations in the same curve. A pilot investigation of the shape of the resilience curve was carried out for the two indicators and showed that the curve could exhibit strong fluctuations, especially with regard to punctuality. These fluctuations were largely canceled out with the composite indicator, which indicates there may be some (currently unknown) interaction between traffic intensity and punctuality. The composite indicator also helps account for the fact that not every disruption has the same impact on the train service. In one case, trains may keep running while delays start to build up, where in another case, some trains may be canceled while the remaining ones are running on time. More specifically, the added benefit of the punctuality component may be observed for partial blockages (which do not necessarily lead to cancellations, but can still cause delays) and for disruptions where it takes a long time before a VSM is applied. It is known from practice that the latter occurs regularly and may have a significant effect on ProRail's KPIs. Studying the traffic intensity alone would underestimate the impact of a disruption in such cases.

Since performance is defined as a weighted sum, one of the indicators could be made more important than the other. A weight $\lambda < 0.5$ puts more emphasis on punctuality, where a weight $\lambda > 0.5$ puts more emphasis on traffic intensity. Given the fluctuating nature of punctuality and based on the premise that it is more important that trains are running at all than that they are running on time, a weight $\lambda > 0.5$ should be favored. Different values for the weight were tested which led to $\lambda = 0.67$ to be chosen as a starting point. At this value for λ, the resilience curve is relatively well behaved while the punctuality component is dominant enough so that changes in punctuality may be observed. Also, $\lambda = 0.67$ is relatively easy to communicate as this makes traffic intensity twice as important as punctuality. The effect of the choice of the weight on the resilience curve is illustrated for one of the studied disruptions in Figure 4.1, which shows the curves for an overhead line failure between The Hague Central and Ypenburg on August 18, 2019. The curve for $\lambda = 0.67$ is shown in gold. Notice the complex interaction between traffic intensity and punctuality, and how the peaks and valleys are smoothed in the golden curve. In this case, punctuality dropped rapidly already before the reported start of the disruption and again during the restart, which would not have been observed if only the traffic intensity was studied. The changes in punctuality can be observed though with the composite indicator.

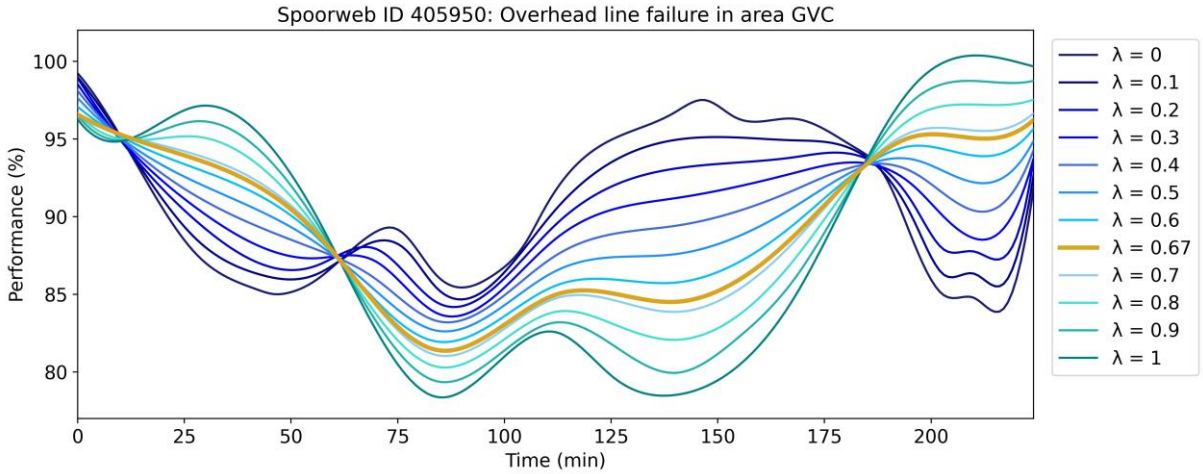


Figure 4.1. Resilience curve for different performance weights for an example disruption.

Resilience metrics

The selection of multiple resilience metrics makes it possible to account for the multidimensional nature of resilience, which could be observed in the resilience curve through different degradation and recovery behaviors. Similar to the selection of performance indicators, the metrics should be complementary and have no or limited overlap to avoid multicollinearity. An overview of the metrics identified in Chapter 2 is presented in Table 4.2. The table briefly describes each of the metrics and the dimension they represent.

Table 4.2. Available resilience metrics per dimension.

Dimension	Metric	Description
Time	**Recovery time**	The total duration of a disruption.
	Recovery rate	The average slope of the resilience curve in the third phase.
	Deterioration rate	The average slope of the resilience curve in the first phase.
Performance	Initial impact	The vertical distance between target performance and minimum performance in the first phase.
	Maximum impact	The vertical distance between target performance and minimum performance.
	Residual functionality	The vertical distance between zero performance and minimum performance.
Time and performance	**Performance loss**	The area enclosed by target performance and the resilience curve.
	ITAE	Performance loss multiplied by a time penalty.
	Profile length	Length of the recovery profile.
	Weighted sum	Time-dependent deviation from a linear recovery profile.

Based on the four metrics printed in bold in Table 4.2, seven resilience metrics were defined. First, the time dimension was captured. Rather than defining recovery time as the total duration of a disruption, such as in Nicholson et al. (2015) or Chan and Schofer (2016), a distinction was made between the resilience phases to account for the unique processes and characteristics of each phase. Thus, *degradation time* gives the duration of the first phase; *response time* gives the duration of the second phase; and *recovery time* gives the duration of the third phase. Second, the performance dimension was captured. *Maximum impact* was selected because it is measured over the entire duration of a disruption, as opposed to the initial impact. Third, the joint dimension of time and performance was captured. *Performance loss* was selected to represent the area above the resilience curve. This metric was preferred over the integral of time absolute error (ITAE), which is mainly useful to describe oscillatory behavior around the optimal performance. However, oscillatory behavior cannot occur for the chosen indicators since optimal performance represents the upper limit of what is theoretically achievable.

To capture time and performance specifically for the transition phases, the weighed sum was considered. Munoz and Dunbar (2015) used it to evaluate supply chain resilience, but it has not been applied in a railway context before. However, during the experiments discussed in Chapter 5 it became apparent that the weighted sum is easily inflated due to the time penalty. This means that large positive or negative values could result from relatively small deviations from a linear degradation or recovery, which makes the outcome of the weighted sum difficult to interpret. Thus, it was decided to abandon the time penalty and only consider the summed deviation from a linear profile. This does not imply that it is assumed that degradation and recovery are linear; it merely provides a reference frame for determining the nonlinearity of the resilience curve in the transition phases. The *degradation profile* represents the summed deviation from a linear degradation, and the *recovery* profile represents the summed deviation from a linear recovery. In contrast with the previous metrics, the degradation and recovery

profile can take on negative values as well. A negative value indicates a concave deviation from a linear profile, where a positive value indicates a convex deviation.

Figure 4.2 shows an enriched resilience curve based on Figure 2.1 including the selected resilience metrics. In this figure and in the definition of the resilience metrics, $Q(t)$ represents the performance function, $f(t)$ represents the linear degradation function and $g(t)$ represents the linear recovery function. The metrics are defined as follows:

1. Degradation time (DT) is defined as the duration of the first resilience phase, from the start of the disruption to the start of the response phase.

$$DT = T_1 - T_0 \tag{3}$$

2. Response time (RST) is defined as the duration of the second resilience phase, from the start of the response phase to the start of the recovery phase.

$$RST = T_2 - T_1 \tag{4}$$

3. Recovery time (RCT) is defined as the duration of the third resilience phase, from the start of the recovery phase to the end of the disruption.

$$RCT = T_3 - T_2 \tag{5}$$

4. Maximum impact (MI) is defined as the vertical distance between target performance Q_0 and minimum performance Q_{min} occurring at any point during the disruption.

$$MI = Q_0 - Q_{min} \tag{6}$$

5. Performance loss (PL) is defined as the area enclosed by target performance Q_0 and the performance function $Q(t)$, calculated as the sum of the area of each interval $[t_i, t_{i+1}]$ where n is the number of intervals i and $T_0 \leq t_i \leq T_3$.

$$PL = \sum_{\substack{1 \leq i \leq n \\ Q(t_i) < Q_0}} \big(Q_0 - Q(t_i)\big)(t_{i+1} - t_i) \tag{7}$$

6. Degradation profile (DP) is defined as the sum of the vertical distance between the linear degradation function $f(t)$ and the performance function $Q(t)$ calculated over the m equally spaced measurement points j in the first phase, where $T_0 \leq t_j \leq T_1$.

$$DP = \sum_{j=1}^{m} \Big(f\big(t_j\big) - Q\big(t_j\big)\Big) \tag{8}$$

7. Recovery profile (RP) is defined as the sum of the vertical distance between the linear recovery function $g(t)$ and the performance function $Q(t)$ calculated over the s equally spaced measurement points k in the third phase, where $T_2 \leq t_k \leq T_3$.

$$RP = \sum_{k=1}^{s} \big(g(t_k) - Q(t_k)\big) \tag{9}$$

For all resilience metrics it holds that a larger positive value indicates worse performance in terms of resilience. Note how Equation (7) has been converted into a left-hand Riemann sum compared to the definition of performance loss in Chapter 2. The reason for this is that the actual performance function $Q(t)$ is not known, so the integral can only be approximated. Given that the resilience curve both decreases and increases over the course of a disruption, a left-hand sum will not likely result in a significant under- or overestimation of the integral. The

sum was defined to include only performance measurements below target performance. This was done to prevent an underestimation or even a negative value of performance loss in case performance is already above target performance when the train service is restarted, which was taken as the starting point for determining the end of disruption as explained in Section 4.2. Also note that Equations (8) and (9) do not approximate an integral but are a summation over points, although the result would be equal to a Riemann sum of the net area between the linear function and the resilience curve in case a unit interval is used.

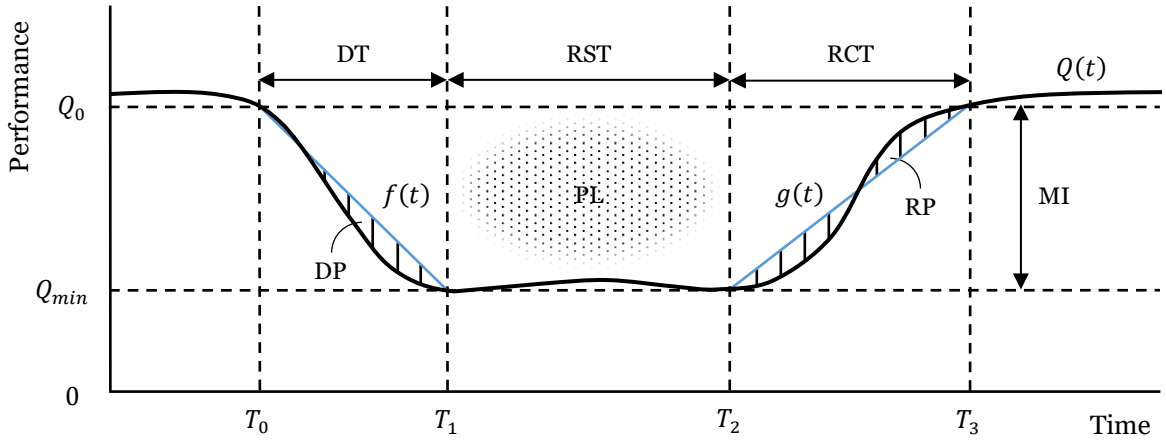


Figure 4.2. Resilience curve including the selected resilience metrics.

4.2. Resilience evaluation framework

The selected performance indicators and resilience metrics were incorporated in the newly developed resilience evaluation framework presented in Figure 4.3, which summarizes the methodology for the data analysis. The framework has been divided into three parts: input, processing and output.

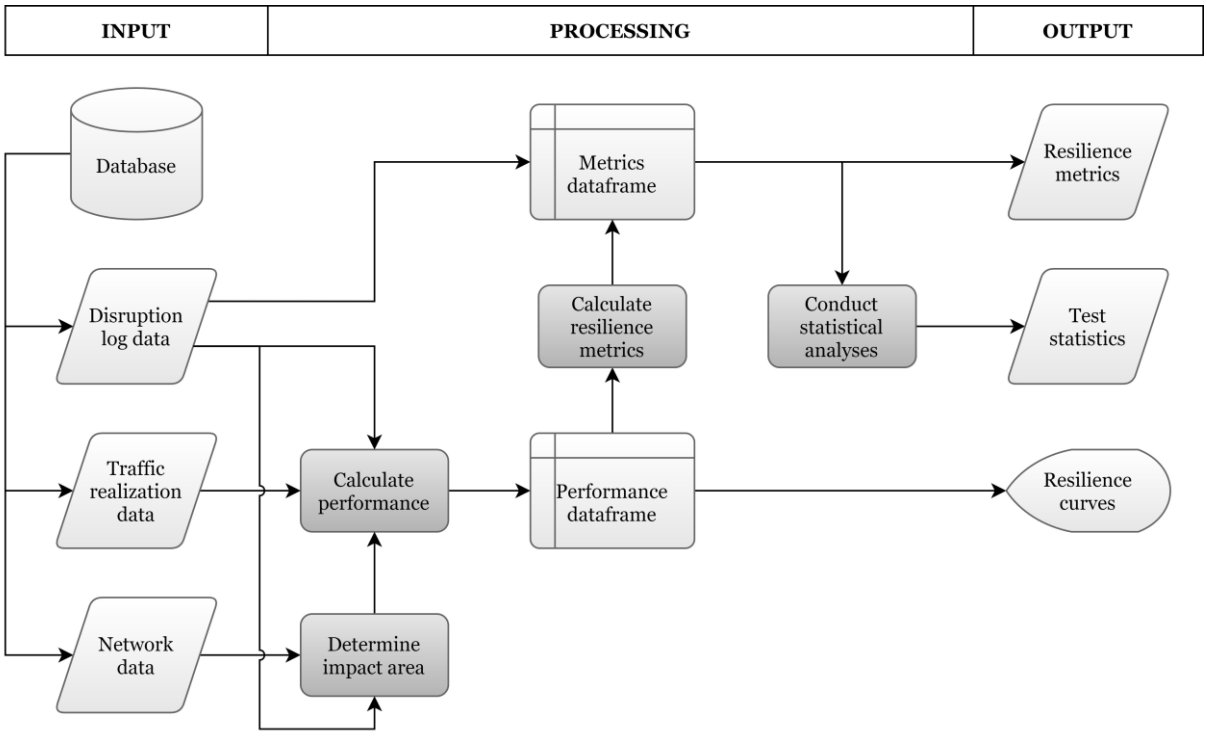


Figure 4.3. The resilience evaluation framework.

In short, the resilience evaluation framework works as follows. The evaluation starts with the collection of traffic realization data, disruption log data and network data. Realization and network data are used to calculate the evolution of performance in the disrupted area according to Equation (2). The performance measurements are then used to calculate the resilience metrics according to Equations (3) to (9). The metrics are stored along with essential information about the disruption retrieved from the disruption log data. Quantitative results are obtained from statistical analysis of the resilience metrics. Graphic results in the form of resilience curves are obtained directly from the performance measurements. The input and processing parts are explained in more detail in this section.

Input

Traffic realization data

The first type of input data is traffic realization data, which can be observed in the activity viewer in Sherlock. The structure of the dataset is long-form: each row corresponds to an observation (in this case: a train activity), and each column corresponds to a variable. The required variables in this dataset are specified by their original column names in Table 4.3. The last four rows of the table describe the various plan times and the realization time of a train activity.

Table 4.3. Overview and description of traffic realization variables.

Variable	Column name	Description
TOC	basic.treinnr_vervoerder	The TOC responsible for the train activity.
Train number	basic.treinnr	The train number corresponding to the train activity.
Service type	basic.treinnr_rijkarakter	The service type corresponding to the train activity.
Timetable point	basic.drp	The timetable point where the activity is registered. Usually the location of a train station.
Control area	basic.drp_post	The traffic control area to which the timetable point belongs.
Activity type	basic.drp_act	The type of activity that is registered. Includes departures, arrivals, passings and shunting movements.
Original VOS time	vklvos.plan_oorspronkelijk	The original plan time in traffic control system VOS.
Actual VOS time	vklvos.plan_actueel	The actual plan time in traffic control system VOS. Usually the same as the original time.
Plan time	basic.plan	The most recent plan time determined from a range of input sources.
Realization time	basic.uitvoer	The most accurate approximation of the realization time.

The exact realization time of a train activity is not known and is therefore approximated using the most accurate source available. This source is usually Trento, a measurement system that determines the realization time of a train activity based on the moment of passing the axle counter or insulated rail joint closest to the particular timetable point. The time gap that corresponds to the distance traveled between the measurement point and the actual location of the timetable point is estimated based on the distance, train length and acceleration or braking behavior of the particular train type. Trento has an accuracy of approximately five seconds. When Trento data are not available, the realization time comes from UIS, which is a component of the traffic control system VOS. Whereas realization times in UIS are measured in real time, realization times in Trento are not. Still, the accuracy of UIS is lower compared to Trento.

Missing realization times must be addressed in data preparation as it could occur that a train activity has no realization time entry. This does not necessarily mean that the activity was canceled. A missing realization time entry can mean one of the following four things:

1. The activity was canceled.
2. No accurate approximation of the realization time is available.
3. The entire path of the particular train was erroneously canceled.
4. The train was included in the year plan, but not in the timetable.

In the last case, the VOS columns can be used to filter out these activities in order to properly identify cancellations. When neither of the columns has a data entry, the activity did not take place and was also never intended to. The only activities that should be counted as cancellations are activities with an original VOS time and plan time, but without an actual VOS time and realization time. Yet, it may occur that an entire train path was canceled in VOS when a train did in fact complete part of its journey. As there is no way of checking when or how often this occurs, it is acknowledged that performance in terms of traffic intensity may be slightly underestimated in some cases.

Disruption log data

The second type of input data is disruption log data, which was exported from the Spoorweb table viewer in Sherlock. Spoorweb is the online platform that enables information sharing and communication among traffic controllers, TOCs, contractors etc. Again, the data structure is long-form, with each row corresponding to a disruption. The viewer contains essential information about disruptions, including many time entries that mark decision points or milestones in the disruption management process. The time entries in the disruption log data are used in ProRail to determine the start and end of the resilience phases. Most entries are logged in real time, but some are approximated in Sherlock. The relevant variables in this dataset are specified by their original column names in Table 4.4.

Table 4.4. Overview and description of disruption log variables.

Variable	Column name	Description
Spoorweb ID	IncidentID	The Spoorweb record ID of the disruption.
Specific cause	IncidentLabel	The specific cause of the disruption as reported by the control room.
Control area	Dvlpost1	The traffic control area in which the disruption has occurred.
Impact type	Logistiek_VDBs	The impact type of the disruption.
Boundary points	Logistiek_VDB_ Begrenzingpunten	The boundary points of the disruption according to the capacity reallocation.
Start of disruption	T_voorval	The moment when the disruption approximately occurred.
First VSM applied	T_gekozeneerstevsm	The moment when the first VSM was applied.
Restart initiated	T_opstartenmogelijk	The moment when the VLC reports that the restart can be initiated.
End of ICB	T_EindeIncidentICB	The moment when the end of the disruption was reported by ICB.
End of disruption	T_EindeIncident	The moment when the end of the disruption was reported by the control room.
Service restored	T_Treindienstopgestart	The moment when two consecutive trains of each train series in the VSM have run again, except for the last series, which should have run only once.
Record closed	T_afsluit	The moment when no further logistic measures are required and the Spoorweb record is closed.

Given the available time entries in the disruption log data, the formal definitions in ProRail of the start and end time of the resilience phases ($T_0 \dots T_3$ as in Figure 4.2) are introduced to allow a comparison to these timepoints later on. First, T_0 is defined as the moment when a disruption occurs, which is approximated as a standard five minutes prior to when the notification reaches the control room. Second, T_1 is defined as the moment when all trains series in the VSM have short-turned for the first time. If no trains have short-turned, this entry is not available. Also, it creates a distorted picture for disruptions that involve long routes* or train series with a low frequency (e.g. one train per hour), since the first short-turns may be performed relatively late. This timepoint may be reported quite early as well (even before the VSM is applied) in case the prescribed short-turns were already performed as emergency measures in the first phase. Most of the interviewed respondents preferred to see T_1 defined as the moment when the VSM is applied, which is also the definition that is followed in this thesis. Third, T_2 is defined as the moment when the VLC reports that the restart can be initiated. When this entry is missing, T_2 is defined as the moment when calamity control (*Incidentenbestrijding, ICB*) reports the end of disruption. ICB is the division of ProRail that resolves a disruption on site, handles collateral damage and manages the aftermath. When this entry is also missing, T_2 is defined as the moment when the control room reports the end of disruption. Note that reporting the "end of disruption" is misleading as this only indicates the end of the second phase. Fourth, T_3 is defined as the moment when the train service has been restored. This moment is approximated in Sherlock as the moment when two consecutive trains of each train series in the VSM have run again according to the regular timetable, except for the last series, which should have run only once. In theory, this might be an appropriate moment to choose for T_3, though the approximation is occasionally missing or at least questionable. For example, a case was studied in which "service restored" was reported at 20:22 when in fact, traffic had already been resumed by the start of the afternoon. The cause of this anomaly was a regional series that had been canceled every other half hour for the rest of the day. In case "service restored" is not available, T_3 is defined as the moment when the Spoorweb record is closed.

Network data

The third type of input data is network data, which specifies how each timetable point is connected to its nearest neighbor(s) in the network. The text file that was retrieved for this purpose is a DONNA infrastructure file from Infra Atlas, which normally serves as input for DONNA, the application used in timetable design. The details of this file, which was accessed through a shared folder, are found in Appendix D. Each row in the file corresponds to a connection to a neighboring timetable point by a single track. By looping though the rows, an adjacency list was created that specifies the neighbor(s) of each point. The list is required to determine the disrupted area for which to draw the resilience curve, as is explained in the next subsection.

Processing

Impact area

For the calculation of system performance from the input data, it is first necessary to determine the area for which the curve is drawn. Ideally, this would be a fixed area such as the entire network, which provides a consistent reference frame for comparing the resilience metrics among disruptions. However, the impact of an individual disruption on the entire network is typically low and the probability that multiple disruptions are observed at the same time is high. The question is therefore which alternative approach could give an accurate representation of the affected area and allow single disruptions (i.e. disruptions that are not connected to any other disruption) to be studied individually. Several resilience curves were therefore compared in a pilot investigation of the disrupted area for a number of disruptions. Curves were drawn for the first and second impact area; the first, second and third impact area; the own traffic control area; and all traffic control areas contributing to the first and second impact area. The curve for the first and second impact area was usually affected the most. The curve that also included

the third impact area was affected less, which indicates that the impact in this area is generally limited and that the concept of decoupling points seems to work in practice as it should. Thus, it was concluded that studying the first and second impact area is the most practical approach for drawing and analyzing the resilience curve. The impact area of a disruption is determined according to the algorithms explained in Section 4.3. An added benefit of impact areas is that they can be used to identify connected disruptions (i.e. disruptions that cannot be considered as single disruptions). Logically, it also makes sense that the location of a disruption lies at or near the center of the studied area.

Performance

With the impact area and the reported start and end of disruption known, it is possible to take a subset of the traffic realization data for which to calculate the evolution of performance over time. As it may be possible that the reported timepoints are not accurate representations of the actual state of the system, the start and end of disruption were determined by checking at which moment performance dropped below and recovered to target performance, respectively. For the end of disruption, target performance was checked no earlier than $t = T_2$ to avoid an underestimation of the disruption length. Target performance was determined as the average performance over a number of arbitrary, relatively quiet days, each of which experienced no more than four disruptions. The disruptions were deliberately included in the calculation of target performance to ensure that target performance is realistic in most cases. For $\lambda = 0.67$, target performance was determined at 97.0%. In some cases though, the system may already be operating below target performance when a disruption occurs. Searching for target performance would then result in an overestimation of the disruption length, which is why a search limit of one hour before the reported start of disruption was introduced. Likewise, a search limit of three hours after the reported end of disruption was introduced to ensure as much as possible that the observed end of disruption is not affected by other, unobserved disturbances (where no VSM was applied). Still, it should be noted that some cases were encountered where the end of disruption seemed to be correctly identified at more than three hours after "service restored" without any sign of interference from other disruptions or disturbances.

For each disruption, the performance Q at time t_i was calculated as a moving average in order to create a relatively smooth curve. A downside of the backward moving average, which only considers past values, is that an artificial time shift of half the interval size is introduced. The time shift can be eliminated by applying a centered moving average (Bashan et al., 2008). The interval size L and step size S were defined similarly for each disruption. This means that at time t_i the interval $\left[t_i - \frac{L}{2}, t_i + \frac{L}{2}\right]$ is considered, at time t_{i+1} the interval $\left[t_i - \frac{L}{2} + S, t_i + \frac{L}{2} + S\right]$, and so on. At each time step, the oldest activities drop out of the interval and newer activities are added to the interval. The activities in each interval were divided into two dataframes: one containing the realized activities (df1) and one containing the canceled activities (df2). As illustrated in the lines of code below, traffic intensity was calculated as the ratio of realized to total activities. Punctuality was calculated as the ratio of punctual to realized activities, but only for activities with a realization time (and thus, a numerical value for the delay). The performance calculation per time interval was coded in Python as follows:

```
intensity   = len(df1) / (len(df1) + len(df2))
punctuality = (df1[df1['delay'].notna()]['delay'] < punc_tol).sum() /
               df1['delay'].notna().sum()
values[T]   = (weight * intensity + (1 - weight) * punctuality) * 100
```

A pilot investigation of different parameter values for the interval size L and step size S showed that $L = 30$ minutes and $S = 1$ minute are appropriate choices to obtain a fairly well-behaved curve. A smaller interval size (e.g. $L = 15$ minutes) would be better able to show the dynamic

nature of the curve, but this would also make the curve more difficult to analyze. After completing the calculations, performance measurements were stored in a wide-form dataframe, which is a dataframe that represents three variables in a spreadsheet-like manner. The times were used as row indices while the Spoorweb IDs were used as column indices.

Resilience metrics
Given the performance measurements for each disruption, the resilience metrics were calculated according to Equations (3) to (9). All metrics, including the durations of the resilience phases, were calculated based on the resilience curve itself and not on the reported timepoints in Sherlock. Therefore, the timepoints T_1 and T_2 had to be derived from the curve as well. This was achieved by applying a steady state detection algorithm as explained in Section 4.3. A long-form dataframe was constructed where each row corresponds to a disruption. The Spoorweb IDs were used as row indices while the resilience metrics and other variables were used as column indices. An overview of the columns in the dataframe is included in Appendix D.

4.3. Algorithms

Two kinds of algorithms were developed in order to determine the impact area and timepoints required for calculating the resilience curve and resilience metrics. The impact area was determined with the help of a graph search algorithm that returns the timetable points in the first and second impact area. The timepoints were determined with the help of a steady state detection algorithm that returns the start and end of the second resilience phase. The algorithms are explained in more detail in this section.

Graph search

In general, a graph search algorithm is used to traverse a graph and find each node (or: vertex) in the graph. Several graph search algorithms have been invented, including the breadth first search (BFS) by Moore (1959). A breadth first search starts from a source node, referred to as the start vertex, and visits any adjacent, unvisited vertices until none are left. The graph is explored one level at a time, and the vertex from which a neighboring vertex is visited is called the “parent”. The basic algorithm is summarized in three steps:

1. Select a start vertex.
2. Visit any adjacent, unvisited vertices, mark them as visited and insert them in a queue.
3. As long as the queue contains elements, extract the first vertex from the queue and start again from this vertex in step 2.

In the context of this thesis, a 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. This approach is also applied in the VGB solver which is currently being developed at ProRail. An example of the first three levels of a breadth first search is presented in Figure 4.4 for a real part of the network. In the example, Gvc is the parent of Gv, Bkh and Laa, Laa is the parent of Gvm, etc.

Certain characteristics of the network structure and the concept of impact areas required imposing constraints on the basic algorithm. First, it had to be specified that the algorithm does not continue its search along a branch of the network when a second decoupling point is reached, as this point marks the end of the second impact area. Second, it had to be specified that the algorithm cannot follow a path that is not driven by any train, for example due to the inability of trains to make sharp turns or the absence of a railway switch at a certain location. Third, the algorithm had to account for the fact that not all vertices may appear in the same path in opposite directions. For example, on the line between Leiden and Haarlem, timetable point Noordwijkerhout is visited when traveling from south to north, but not when traveling from north to south. Fourth, the algorithm had to allow a vertex to have more than one parent,

and to visit this vertex when it had already been visited from another parent. For example, when traveling from east to west, Amsterdam Sloterdijk (Ass) can be visited from Transformatorweg Aansluiting (Asdta) and Overbrakerpolder Aansluiting (Obpa). When it is visited from Asdta, the path continues towards Schiphol. When it is visited from Obpa, the path continues towards Haarlem. If only one parent was allowed, either Haarlem or Schiphol would never be reached depending on the parent from which Ass is visited first. Fifth, the algorithm had to be able to handle more than one start vertex, since a single start vertex only occurs for a timetable point outage. A line blockage will always have two or even three start vertices.

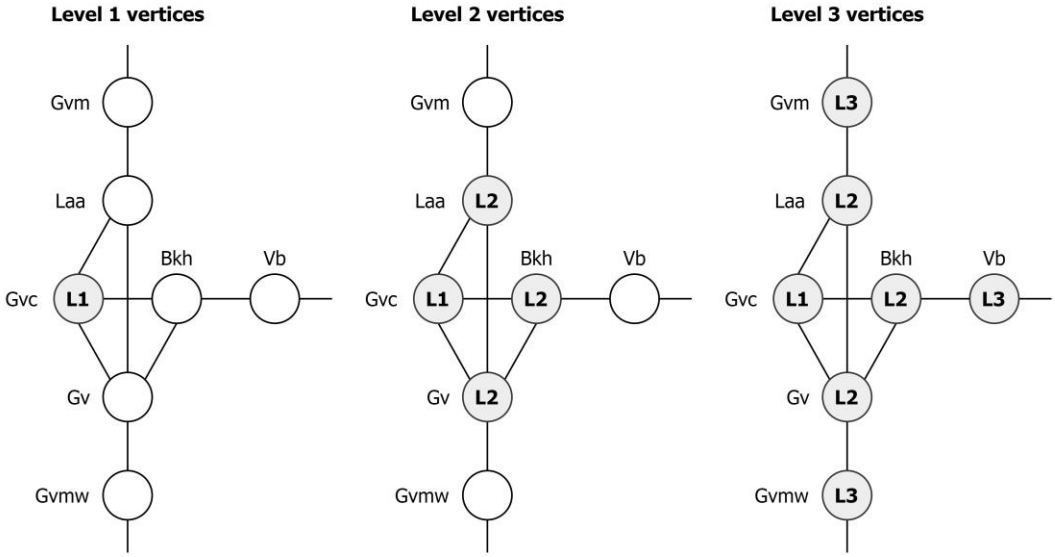


Figure 4.4. First three levels of a breadth first search starting from The Hague Central.

Because of the increasing complexity of the search with each additional start vertex, three separate algorithms were developed for one, two and three start vertices. The algorithms are included in Appendix H. In each of the algorithms, the first four constrains were incorporated as follows. The algorithm knows when to stop searching along a branch of the network by tracking the impact area rather than the level. It checks for feasible paths by comparing a sequence of three vertices (the current vertex n, its parent v and the parent of v) to a concatenated string of all realized train paths in a day. The reverse sequence is checked as well. Multiple parents per vertex are allowed by storing them in a list and reading the parent of a vertex either as a string or a list. In the algorithm for two start vertices, the ability to handle more than one start vertex was incorporated by starting the search from one of the boundary points and breaking when the other one is reached. Then, the algorithm backtracks to the first boundary point to find the first impact area (or part of the first impact area, in case the boundary points are not decoupling points) and continues searching outwards from both boundary points. A similar approach was taken in the algorithm for three start vertices, but here it must be checked how the boundary points are connected. Possible scenarios are:

- The three boundary points lie on the same line (Figure 4.5a).
- The three boundary points do not lie on the same line.
    - All boundary points can be reached from any boundary point (Figure 4.5b).
        - All paths are accurate.
        - One of the paths is a detour; the longest path should be ignored.
    - All boundary points can be reached from only one boundary point (Figure 4.5c).

A large part of the algorithm involves checking which of these conditions is true, and consequently, which area to consider as part of the first impact area. Once this area is determined, the search continues in a similar fashion as for one and two start vertices.

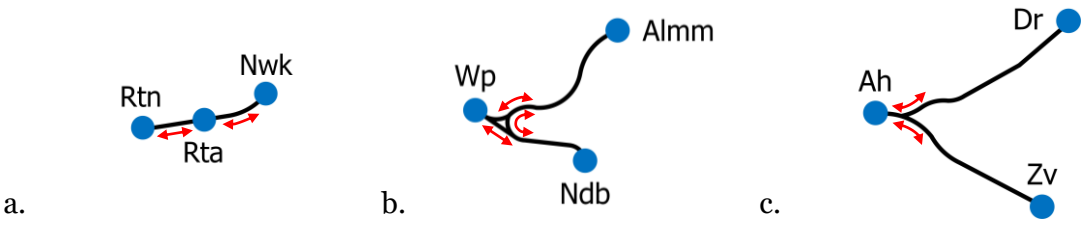


Figure 4.5. Possible scenarios for line blockages with three boundary points.

Steady state detection

In general, a steady state detection (SSD) algorithm is used to identify the steady parts of time series data. Several of these algorithms have been invented, for example by Cao and Rhinehart (1995), Castiglioni and Di Rienzo (2004), Luo et al. (2009), Kelly and Hedengren (2013), Liao et al. (2016) and Dalheim and Steen (2020). Dalheim and Steen (2020) also reviewed a number of current techniques and concluded that many require the definition and finetuning of several filters or model parameters, which can make them difficult to implement. Dalheim and Steen (2020) therefore developed a robust and computationally efficient technique of their own that uses a sliding window. The basic approach of Dalheim and Steen (2020) is as follows. A linear regression model is fitted to the data in each time window with length n. Two consecutive windows overlap by $(n - 1)$ measurement points, as illustrated in Figure 4.6. If the slope of the regression model is significantly different from zero, the window is considered unsteady. This is tested by comparing the t-value from the regression analysis to the critical t-value. The t-value for each window is written to the first and last position of the window in a dataframe. The front of the window marks the change from a steady to an unsteady state if such a change were to occur, while the rear marks the change from an unsteady to a steady state. Only two parameters need to be determined: the significance level α and the window size n. This makes the approach easy to implement.

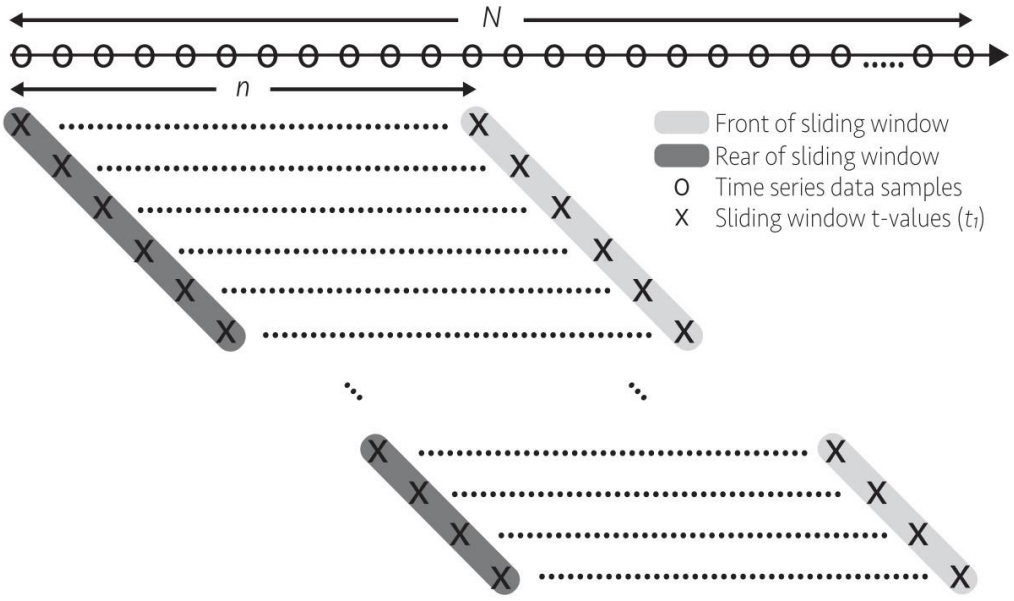


Figure 4.6. The sliding window approach (Dalheim & Steen, 2020).

In this thesis, an adapted steady state detection algorithm is proposed based on the approach of Dalheim and Steen (2020) with the purpose of identifying the second resilience phase. The approach was tested for several dozen disruptions, but only led to a satisfactory result in part of the cases. This is attributed to the diverse behavior of the resilience curve and the inherently different nature of the data compared to Dalheim and Steen (2020), who evaluated ship perfor-

mance and navigation data. These kind of data are generally steady, whereas the resilience curve is more dynamic. Therefore, modifications were made to the algorithm which suit the data better and which enable the algorithm to handle both short and long disruptions as well as calm and noisy curves. The modifications are as follows:

- The window length was made dependent on the total number of measurement points, which is equal to the disruption length. To avoid too long time windows for relatively long disruptions, the square root of the number of measurement points was taken and multiplied by a factor 1.8. A pilot investigation of the window length showed that this works well for both short and long disruptions. A factor 1.5 occasionally resulted in too short windows, whereas a factor 2 occasionally resulted in too long windows.
- The search for a steady state was allowed only when the first measurement point in a time window is below a performance threshold. This is to avoid the detection of a steady state in the first or third phase. A pilot investigation showed that a threshold of 40% is usually adequate to capture the steady state that would be expected by visual inspection of the curve. This means a steady state can be observed when the first measurement in a window is below $Q_{min} + 0.40(Q_{max} - Q_{min})$, where Q_{min} and Q_{max} are the minimum and maximum performance for the studied resilience curve, respectively.
- Evaluation by the front and rear of the sliding window was dropped since the algorithm should only return the first and last steady point in the lower part of the curve. For this application it is not important to know exactly which part of the second phase is steady and which part is not, although this could be a direction for future research.

Apart from these modification, the algorithm compares the p-value of the slope of the fitted line to the significance level instead of comparing the t-value to a critical t-value. A window is marked as steady when the p-value is greater than α. Because it should be really evident when a window is unsteady, a 99% confidence interval was defined. This translates to a significance level of $\alpha = 0.01$. A larger α would mean that fewer time windows are marked as steady. The steady state detection algorithm is included in Appendix I.

4.4. Statistical methods

A range of methods are available for statistical analysis of the resilience metrics to evaluate the differences and similarities in the metrics among disruptions. Two classes of analyses were explored: group comparisons and regression analysis. The different options for these classes of analyses are discussed in more detail in this section.

Group comparisons

Group comparisons help identify the differences among groups based on one or more dependent variables. The independent categorical variable, also referred to as the factor, defines group membership. In the context of this thesis, the dependent variables are the resilience metrics and the factor is the disruption cause. The parametric option for group comparisons is analysis of variance (ANOVA). Nonparametric options are the Kruskal-Wallis test and Welch's ANOVA.

Parametric group comparison

ANOVA evaluates the differences in a dependent variable among more than two groups. It does so by comparing the variance between groups (MS_B) to the variance within groups (MS_W). ANOVA takes both the mean and the variance of the observations into account. The resulting test statistic is referred to as the F-statistic, which is calculated according to Equation (10).

$$F = \frac{MS_W}{MS_B} \tag{10}$$

The F-statistic "compares the amount of variance that can be explained by group membership to the amount of variance that cannot be explained by the group" (Mertens et al., 2017). The null hypothesis is that there are no differences among groups, which is indicated by a low F-statistic. If the null hypothesis is rejected, it means that differences among groups exist, and that splitting the total sample into groups helps clarify the data. As only one categorical variable is used to define group membership, this type of analysis is also referred to as one-way ANOVA. Since ANOVA is a parametric method, it assumes that the data approximately follow a normal distribution. Parametric assumptions must be met to correctly interpret the test results. The assumptions vary with the type of analysis. For one-way ANOVA, they are:

- Independence of observations
- Normality of the dependent variable
- Homogeneity of variance of the dependent variable

ANOVA is an omnibus test which is two-sided by definition. This means it only reveals whether a difference exists among groups. It does not tell where exactly the difference lies or how large it is (Mertens et al., 2017). Therefore, an additional post hoc test is required to explore the results in more detail. Common post hoc tests for ANOVA include:

- Fisher's Least Significant Difference (LSD): Applies a series of pairwise comparisons among groups while controlling for the individual error rate.
- Tukey's Honest Significant Difference (HSD): Applies a series of pairwise comparisons among groups while controlling for the family error rate. This decreases the probability of making a Type I error (also known as a false positive) compared to Fisher's LSD.

Nonparametric group comparison

In case parametric assumptions are violated, an alternative test is available: the Kruskal-Wallis test. This test is considered to be the nonparametric equivalent of one-way ANOVA. Instead of evaluating the means of all samples, Kruskal-Wallis evaluates the medians. The resulting test statistic is referred to as the H-statistic. The Kruskal-Wallis test applies ranking, which means that the numerical value of each observation is replaced by its rank. For example, three observations $\{40, 60, 50\}$ would be ranked as $\{1, 3, 2\}$. A tie occurs when two observations with exactly the same numerical value compete for the same rank. In case the data contain no ties, the H-statistic is calculated according to Equation (11).

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \left(\bar{r}_i - \frac{N+1}{2} \right)^2 \tag{11}$$

Where:

- N is the total number of observations
- g is the number of groups
- n_i is the number of observations in group i
- $\bar{r}_i$ is the average rank of all observations in group i

Although parametric assumptions do not need to be met in order to perform a Kruskal-Wallis test, the test does require independence of observations and identically shaped distributions (and thus, equal variances). Like ANOVA, Kruskal-Wallis is an omnibus test which follows the same hypotheses. This means that when the null hypothesis is rejected, the result is significant and a post hoc test is required to study the observed differences in detail. Common post hoc tests for Kruskal-Wallis include:

- Nemenyi test: Similar to Tukey's HSD, except that it tests between rank means instead of numerical means. This test relies on the studentized range distribution.
- Dunn's test: Similar to the Nemenyi test, except that it relies on the normal distribution and includes a correction for ties.
- Conover test: Similar to Dunn's test, except that it relies on the Student's t-distribution and assumes a different definition of the standard error.

When the assumption of identically shaped distributions is violated, the Kruskal-Wallis test may give inaccurate results. As an alternative, Welch's ANOVA can be considered. This method is similar to one-way ANOVA and applies weights to adjust the grand mean (i.e. the mean of the total sample) based on the group means. Due to this modification, it does not require homogeneity of variance. Since ANOVA is pretty robust against violation of the normality assumption (Mertens et al., 2017), Welch's ANOVA is useful for analyzing data that are nonnormally distributed and have unequal variances among groups. As in regular ANOVA, the resulting test statistic is referred to as the F-statistic, which is calculated according to Equation (12).

$$F = \frac{SS/(g-1)}{1 + \frac{2\Lambda(g-2)}{3}} \tag{12}$$

Where:

- SS is the weighted sum of squares
- Λ is a factor based on the weights and group sizes

The common post hoc test for Welch's ANOVA is the Games-Howell test, which is similar to Tukey's HSD but does not require equal variances. For more details on the calculation of the test statistics for the described methods, the interested reader may refer to Liu (2015).

Regression analysis

Regression analysis helps identify the relationships between one or more independent variables and a dependent variable. With regression analysis, it could for instance be determined if the value of one resilience metric can be inferred from the value of another resilience metric or from other explanatory variables. The size and direction of the relationships offer insights into the dynamics between the resilience phases. This is something the group comparisons cannot provide, because the metrics are studied independently in those comparisons. Just as for group comparisons, there is a distinction between parametric and nonparametric regression models.

Parametric (linear) regression analysis

Arguably the simplest form of regression analysis is linear regression, which assumes that the dependent variable can be predicted by a linear predictor function. When the relationship to a single independent variable is studied, linear regression is referred to as simple regression. The general form of the predictor function is given by Equation (13).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{13}$$

Where:

- Y is the dependent variable
- $X_1 \ldots X_n$ are the independent variables
- β_0 is the intercept at the vertical axis
- $\beta_1 \ldots \beta_n$ are the coefficients assigned to each variable
- ε is the error term which captures all factors not explained by the model

The approach to regression modeling is to fit a line through the measurement points so the error term is reduced as much as possible. This is usually performed by applying the Ordinary Least Squares (OLS) estimator, which minimizes the sum of the squared vertical distance between the fitted line and the measurement points. The resulting betas describe by how much the dependent variable changes for a one unit increase or decrease in an independent variable. In the context of this thesis, a larger value for each of the resilience metrics indicates a less resilient system, and so, improvement efforts may be focused on the independent variable with the largest positive beta. Like ANOVA, linear regression is a parametric method, which means parametric assumptions must be met. For regression analysis, the assumptions are:

- Independence of residual errors
- Normality of residual errors
- Absence of multicollinearity
- Homoscedasticity of residual errors
- Linearity

According to the Gauss-Markov theorem though, the first two assumptions do not have to be met as long as the errors are uncorrelated and have a mean expected value of zero. In that case, OLS is still the best linear unbiased estimator (Theil, 1971). In case the linearity assumption is violated, one can either choose to transform the data to a linear form (for example by taking the logarithm) or resort to nonlinear regression analysis.

Nonparametric (robust) regression analysis
When parametric assumptions are violated and transforming the data or expanding the model does not help, robust analysis methods offer an alternative. Robust methods are designed to be sufficiently insensitive to deviations from the parametric assumptions (Draper, 1988). A particular advantage of robust regression models is that they are less sensitive to outliers. Well-known examples of robust regression models include the following:

- Huber regression: Assigns less weight to observations with large residuals.
- Theil-Sen regression: Fits multiple lines to pairs of observations, then takes the median of the slopes of all lines.
- Random Sample Consensus (RANSAC): Separates the dataset into outliers and inliers, then fits a line to the set of inliers.

4.5. Chapter summary

This chapter started with the selection of performance indicators and resilience metrics for a quantitative description of the resilience curve. The resilience evaluation framework was presented and the input data and calculation procedure were discussed. A graph search algorithm and a steady state detection algorithm were presented which can identify the impact area and the steady state in the resilience curve, respectively. Lastly, statistical methods were discussed by explaining the basic principles of the available parametric and nonparametric options for group comparisons and regression analysis.

Answer to subquestion 3

With the knowledge obtained in this chapter, the third subquestion is answered.

> **Subquestion 3:** How can the spatiotemporal effects of disruptions and recovery measures on railway system performance be quantified for the different resilience phases?

When defining system performance, it is important that no two indicators represent the same functionality. Two indicators were preferred over the others: traffic punctuality, which represents the functionality of travel time, and traffic intensity, which represents the functionality

of use of resources. Punctuality and traffic intensity are complementary, since punctuality does not consider cancellations and traffic intensity does not consider delays. The two indicators were combined in a composite performance indicator which is calculated as a weighted sum. In the weighted sum, more weight should preferably be put on the traffic intensity component because of 1) the potentially strong fluctuations in the punctuality component, and 2) the premise that it is more important that trains are running than that they are running on time.

The spatiotemporal effects of a disruption may be quantified and visualized in the resilience curve in terms of the selected performance indicators. The temporal effects were incorporated by calculating performance as a centered moving average over an interval of 30 minutes. The spatial effects were incorporated by calculating performance for all timetable points in the first and second impact area. As performance is calculated relative to the timetable and is expressed in percentages, it is not biased by the size of the area. Still, when the spatial impact is substantial, more points in the area will be affected and performance will be lower as a result.

Seven resilience metrics were defined to describe the profile of the resilience curve quantitatively. The degradation time, response time and recovery time were selected to represent the time dimension. Maximum impact was selected to represent the performance dimension. Lastly, the performance loss, degradation profile and recovery profile were selected to represent the combined dimension of time and performance. Together, these seven metrics should account for the multidimensional nature of resilience.

5. Case study and results

This chapter presents the results of the data analysis based on the resilience evaluation framework. Results were obtained from a number of experiments with a focus on single disruptions. Connected disruptions were briefly explored as well. The goal of the experiments was to incrementally gain a better understanding of the behavior of the resilience curve, the resilience metrics, and the dynamics between the resilience phases for different types of disruptions. Section 5.1 introduces the case study and presents an overview of the experiments. Section 5.2 presents the resilience curve, steady state, timepoints, resilience metrics and impact area for an example disruption. Section 5.3 presents representative resilience curves for different types of disruptions and discusses the differences between the observed and reported timepoints. Section 5.4 presents the results of the group comparisons which aimed to identify differences between the resilience metrics among disruption causes. Section 5.5 presents the results of the regression analyses which aimed to identify relationships between the resilience metrics and with other explanatory variables. Section 5.6 discusses the resilience curve for two separate connected disruptions. Section 5.7 presents the networkwide resilience curves for a number of red and black days. Section 5.8 summarizes the chapter and provides the answer to subquestion four.

5.1. Case description

The resilience evaluation framework was applied in a case study on the Dutch railway network. This section covers the study area and period; the studied disruption causes; the identification of connected disruptions; handling missing, invalid and inconsistent data; and the order and contents of the experiments.

Study area and period

The study area includes the entire Dutch railway network with the exception of traffic control area Kijfhoek. Because Kijfhoek exclusively handles freight traffic, it cannot be treated in the same way as the rest of the network which handles both passenger and freight traffic. Thus, the study area contains 684 timetable points spread across 12 traffic control areas. Studying disruptions in the entire network has two main benefits. First, it helps identify as many potentially different resilience curves as possible. Second, the greater the amount of data, the more the sensitivity to outliers is reduced. The study period was limited to timetable year 2019, which was the last regular timetable year before the COVID-19 pandemic. ProRail considers this as the benchmark for returning to the regular timetable when COVID-19 restrictions are lifted. In total, 2,152 disruptions with a logistical record were observed in this period. Part of the disruptions occurred on the six black days and three near-black days presented in Table 5.1. The table specifies the date, cause and networkwide performance for each day. The black and red day of November 27 and 28, respectively, were difficult to attribute to a single cause (ProRail, 2020), which is why these were labeled as “multiple causes”. Because of the extreme conditions on these nine days, they were excluded from the majority of the experiments.

The collected traffic realization data for the study area and period include all possible service and activity types. Not all of these data were relevant for the experiments. The studied service types and their abbreviations in the data include regional rail (SPR, ST, S), intercity rail (IC), high-speed rail (HSN) and international rail (THA, ICE, ES). The studied activity types include arrivals, short stops (which are arrivals planned in the same minute as the departure) and passings. Counting the departures as well would mean that a train is observed twice at the same location when it makes a stop, which results in double counting. Yet, excluding the departures does not mean that departure delays are overlooked: when a train is delayed during a stop at the platform, for example due to a longer dwell time, the departure delay is simply observed at the next activity. After filtering the realization data for the specified activities, the datafile contained 49,005,094 rows, which amounts to approximately 132,000 activities per day.

Table 5.1. Red and black days initially excluded from the experiments.

Date	Type of day	Cause	Punctuality (%)	Cancellations (%)
22-01-2019	Black	Snowfall	77.8	20.0
30-01-2019	Black	Snowfall	89.3	18.4
18-03-2019	Black	Regional strike	89.9	13.8
28-05-2019	Black	National strike	95.2	95.2
24-06-2019	Red	Collision in Utrecht	79.2	9.8
25-07-2019	Red	Extreme heat	83.2	9.6
26-07-2019	Black	Extreme heat	87.4	11.8
27-11-2019	Black	Multiple causes	74.2	5.5
28-11-2019	Red	Multiple causes	75.7	3.7

Studied disruption causes

Disruptions that match the top five specific causes were analyzed as single disruptions in this case study. This includes train defects, section or signal failures, collisions, switch failures and overhead line failures. Together, these disruptions made up 76% of all disruptions requiring a capacity reallocation in 2019. Section and signal failures were considered as a single cause because they are often related. The various disruption causes are defined as follows:

- A train defect is a defect in the train itself such as a brake failure, electronic failure, fire, smoke development, improperly closing doors or a damaged pantograph.
- A section failure is a situation where there is a problem with the train detection on a track section, which usually means that the section is falsely reported as occupied.
- A signal failure is a situation where a signal reverts to a fail-safe state and shows a red signal aspect, for example due to a disrupted power supply or a section failure.
- A collision is an encounter of a train with an obstacle such as a person, animal, road vehicle or infrastructure object.
- A switch failure is a railway switch that can no longer be fixed in the right position, for example due to a mechanical failure or obstruction by snow, twigs or other objects.
- An overhead line failure is the loss of power in an overhead line group, for example due to a broken cable or an object that is stuck in the overhead lines.

Identification of connected disruptions

Although the focus of this case study was placed on single disruptions, in practice it may occur that simultaneous disruptions affect each other which may require VSMs to be adjusted. These disruptions are “connected”. Note that this does not necessarily mean that they share a causal relationship. However, such disruptions should not be studied independently from each other, since the impact of one disruption may contaminate the resilience curve of another disruption. Zhu and Goverde (2021) defined connected disruptions as two or more disruptions that:

1. Have overlapping time periods;
2. May start or end at a different time;
3. Occur at different geographic locations;
4. Are connected by one or more train series.

Tracing back all the train series involved in a disruption would be a cumbersome task. Instead, in this case study the impact areas were evaluated. It was stated that, in addition to the first three conditions, disruptions are connected if they have at least one timetable point in their impact area in common. This approach is not only more practical considering that the impact area already had to be determined for each disruption, it is also more appropriate for handling train series that run on long routes. For example, it prevents a disruption around Amsterdam from being connected to a disruption in the south of the Netherlands. In practice, those disrup-

tions would not be connected as there are plenty of decoupling points in between. Overlapping time periods were checked for the reported start and end times before calculating the resilience curves. As disruptions were frequently observed to last longer in reality than reported, a second check was performed for the observed start and end times after calculating the curves. While this does not guarantee that there is absolutely no interaction between any two single disruptions, the risk of contamination was reduced as much as reasonably achievable this way.

Handling missing, invalid and inconsistent data

The disruption log data (and to a lesser extent, the traffic realization data) rely on human input. Therefore it was necessary to handle missing, invalid and inconsistent data entries. In the realization data only the rows matching the specified TOCs, service types and activity types were preserved. It was specified that these rows should always have a valid train number, timetable point and plan time. Handling missing, invalid and inconsistent entries in the disruption log data was more complicated. Several filters were applied in consecutive filtering steps. Filtering the data in steps helps create a better understanding of which part of the data is lost. An overview of the remaining number of disruptions per cause, impact type and control area after each filtering step is included in Appendix J. In all filtering steps combined, the following cases were removed to find the set of disruptions that could potentially be studied as single disruptions:

- Disruptions on extreme days.
- Disruptions in traffic control area Kijfhoek.
- Disruptions that do not match the top five causes.
- Disruptions with a connection to one or more other disruptions.
- Disruptions with an impact area of less than six timetable points[3].
- Disruptions with a reported duration longer than ten hours[4].
- Disruptions with a missing time entry for “restart initiated”.

In addition, five cases were removed manually. This concerns the disruptions with IDs 359363, 359402, 359406 and 359412, which were all affected by the control center outage in Maastricht in the afternoon of March 10, 2019. The fifth case that was removed is the section failure near Zutphen on April 7, 2019 with ID 367232. The disruption caused a line blockage with not three, but four boundary points. Five other cases with four boundary points were observed, but those could be converted to a two or three boundary point disruption. Instead of developing a dedicated algorithm for the one remaining case with four boundary points, it was removed from the data after verifying that this disruption could not be connected to any other disruption.

Some results of the data filtering process are worth discussing. First, 51% of the cases were lost in the transition from disruptions matching the top five causes towards single disruptions, which means that 791 disruptions had one or more connections and 750 disruptions had none based on the reported timepoints. Most connected disruptions appeared in the traffic control areas in the Randstad (Amsterdam, Amersfoort, Utrecht, The Hague and Rotterdam), which is explained by the relatively high network density in this area. Second, particularly many switch failures were lost in the transition towards single disruptions. Only 50 of the 153 switch failures (33%) could be classified as single disruptions. Studying the connected switch failures in detail revealed that most connections relate to train defects (33%), which is explained by the fact that

---

[3] Three possible impact areas were observed that contain only three timetable points: the two boundary points and one point in between. Because of a lack of passenger traffic in those areas, the breadth first search algorithm did not search beyond the boundary points. The smallest valid impact areas that were observed contain six timetable points.

[4] For disruptions longer than approximately ten hours, checking the performance threshold after the reported end of the second phase becomes problematic because it results in a significant overestimation of the actual disruption length.

train defects are simply the most common. In second place, switch failures were connected to other switch failures (24%). A telling example is the morning of December 16, 2018 when seven switch failures were reported at different locations around Amsterdam due to snowfall. In the end, 706 disruptions remained which could potentially be analyzed as single disruptions. An overview of the number of disruptions per cause and impact type is presented in Table 5.2.

Table 5.2. Number of potential single disruptions per cause and per impact type.

Specific cause	Number	Impact type	Number
Train defect	346	Full timetable point outage	10
Section/signal failure	141	Partial timetable point outage	45
Collision	146	Full line blockage	323
Switch failure	47	Partial line blockage	299
Overhead line failure	26	Reduced timetable point functionality	17
		Reduced line functionality	12
Total	706	**Total**	706

In the experiments, only those disruptions were analyzed for which the resilience curve could be described properly, in other words: for which calculating the resilience metrics was justified. For example, calculating the recovery time or recovery profile would have no meaning if the end of disruption cannot be identified. Therefore, the following cases were excluded:

- Disruptions for which the curve remained above target performance from the reported start until end of disruption.
- Disruptions for which the start time could not be identified from the realization data, which means the curve was already below target performance at $T_{0,S} - 60$ minutes.
- Disruptions for which the end time could not be identified from the realization data, which means the curve did not yet recover to target performance at $T_{3,S} + 180$ minutes.
- Disruptions that were connected based on the observed start and end time.
- Disruptions for which an empty time window was encountered.
- Disruptions for which no steady state could be identified.

For $\lambda = 0.67$, the remaining number of disruptions was 445 out of 706. Similar results were obtained for other values of λ. This means that from the original set of 1,541 disruptions, less than one third could eventually be analyzed as single disruptions.

Experiments

Characteristics of the resilience curve for an example case

Experiments started with the detailed evaluation of one arbitrary disruption. The aim of this evaluation was to build trust in the followed methodology and illustrate how the algorithms are brought to practice. For this disruption, the resilience curve was drawn for different performance weights; the profile of the curve was explained based on the logging in Sherlock; the steady state and the observed and reported timepoints were examined; the resilience metrics were calculated; and the impact area was plotted in a network diagram.

Representative resilience curves and time differences

In the next part of the experiments, representative resilience curves were drawn with the aim to create an understanding of the different resilience curve behaviors that may be observed. First, distinctive types of resilience curves (such as the bathtub) were identified. Second, the mean resilience curve was drawn per disruption cause for $\lambda = 0.67$ and for the separate contributions of punctuality ($\lambda = 0$) and traffic intensity ($\lambda = 1$). Third, the differences between the observed timepoints $T_0 \dots T_3$ and the reported timepoints $T_{0,S} \dots T_{3,S}$ (with a subscript "S" for Sherlock) were evaluated to identify discrepancies in the timepoints.

Comparison of resilience metrics across disruption causes
As part of the statistical analyses, group comparisons and post hoc tests were performed to identify differences and similarities in the resilience metrics. Disruptions were grouped by their specific cause. This means there were five groups and seven comparisons to be made: one for each metric. The comparisons were initially performed for $\lambda = 0.67$ and verified for other values of λ in the range of 0.5 to 1.

Relationships between resilience metrics
In subsequent statistical analyses, regression models were fitted to identify relationships between the resilience metrics and other explanatory variables. If such relationships exist, it would be worth conducting follow-up research on the underlying causes or even monitoring the metrics in real time. Again, the analyses were initially performed for $\lambda = 0.67$ and verified for other values of λ in the range of 0.5 to 1.

Resilience curves for connected disruptions
Following the experiments regarding single disruptions, resilience curves were drawn for connected disruptions as well. For two arbitrary cases, the resilience curve was drawn for the total duration and combined impact area of the individual disruptions that constitute the connected disruption. The impact area was shown as well. Characteristics of connected disruptions were identified which deserve attention if one were to evaluate connected disruptions in more detail.

Networkwide resilience curves for extreme days
In the last experiment, the resilience curves were drawn for the previously excluded extreme days. As it is assumed that disruption effects propagate more strongly through the network on those days, it might be hard to distinguish between single or even connected disruptions. For this reason, the curves were drawn for the entire day on a networkwide scale as in Dekker et al. (2021) and compared to the networkwide resilience curve for a number of regular days.

5.2. Characteristics of the resilience curve for an example case

This section presents a detailed evaluation of an arbitrary disruption, which illustrates the working of the resilience evaluation framework. For this disruption, the resilience curve was drawn for the first and second impact area; the effect of different performance weights was investigated; the steady state and timepoints were identified; the resilience metrics were calculated; and the impact area was shown in a network diagram.

The example case is a collision that occurred in traffic control area Amersfoort between Putten and Nunspeet on April 4, 2019. The collision occurred at 14:43 and resulted in a full line blockage of the double track line. The first impact area (bounded by Amersfoort and Zwolle) and the second impact area (bounded by Steenwijk, Assen, Almelo, Deventer, Hilversum and Utrecht) comprised a total impact area of 64 timetable points, which is larger than the average traffic control area. The resilience curve for this disruption was drawn for different performance weights. The curves are presented in Figure 5.1. It is observed that traffic intensity ($\lambda = 1$) dropped quickly at the start of the disruption. At the lowest point, traffic intensity measured 79.3%. In contrast, punctuality ($\lambda = 0$) remained relatively stable around 96% until approximately 100 minutes into the disruption. This illustrates how the rescheduling process is meant to work: in the beginning, enough trains were canceled or short-turned so the remaining trains did not experience delays, thereby preventing cascading effects. The moment when punctuality eventually dropped matches the moment when evacuation of the stranded passengers and retrieval of the damaged train began. During this time, punctuality dropped to 86.1% at the lowest point, while traffic intensity had already partly recovered. All resilience curves recover similarly towards the end of the disruption, which indicates that the restart was executed well and did not cause many new delays. Assessing the curves from start to end, 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. Thus, $\lambda = 0.67$ is still considered an appropriate starting point for use in the composite indicator.

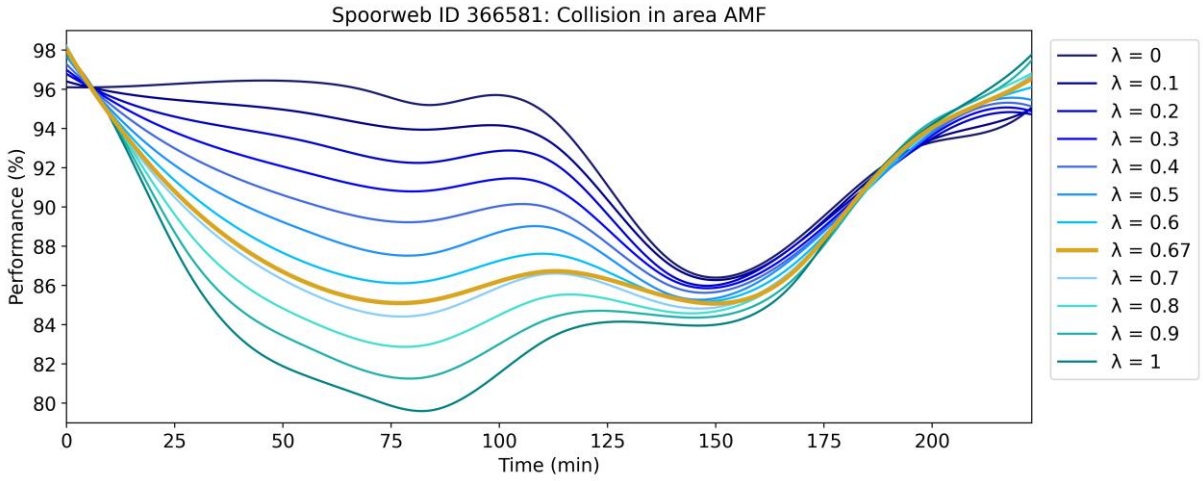


Figure 5.1. Resilience curve for different performance weights for the studied disruption.

The steady state detection algorithm was applied to the resilience curve for $\lambda = 0.67$, which resulted in the timepoints T_1 and T_2. The timepoints T_0 and T_3 were derived directly from the performance calculation. The reported timepoints $T_{0,S} \dots T_{3,S}$ were obtained from the disruption log data. The steady state and the observed and reported timepoints are presented in Figure 5.2, which shows that a data-driven approach to determine the timepoints is rightly preferred over retrieving the timepoints from the disruption log data. The steady parts of the curve are shown in green, where the unsteady parts are shown in red. The detection of a steady state in the curve was successful, since it matches the steady state that one would identify by observation and it is not affected by the slight change in performance during the second phase. Regarding the timepoints, is observed that the disruption was reported at $T_{0,S} = 7$ minutes relative to the moment when the curve dropped below target performance. The VSM was applied at $T_{1,S} = 25$ minutes, just 18 minutes after $T_{0,S}$. Yet, the steady state was not reached until $T_1 = 68$ minutes. The restart was initiated at $T_{2,S} = 163$ minutes, which coincides with the observed timepoint T_2. According to the approximation in Sherlock, the train service was restored at $T_{3,S} = 210$ minutes, which is slightly earlier than the observed end of the disruption, $T_3 = 223$ minutes. In total, the disruption was found to last 20 minutes longer than reported.

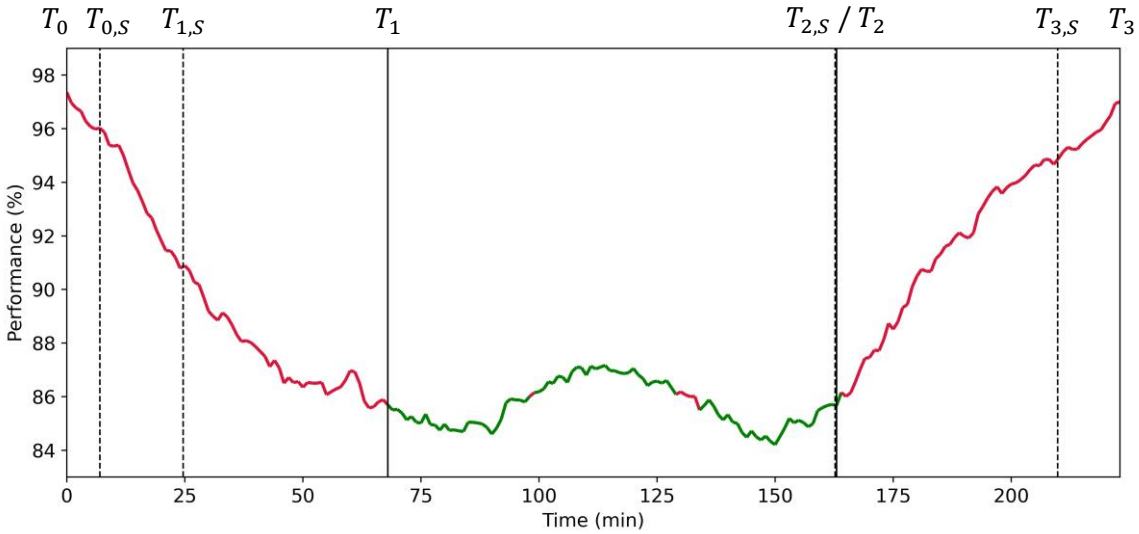


Figure 5.2. Steady state and timepoints for the studied disruption.

Based on the resilience curve in Figure 5.2, the resilience metrics were calculated as follows:

- $DT = 68.00$ minutes
- $RST = 95.00$ minutes
- $RCT = 60.00$ minutes
- $MI = 12.80$ percentage points
- $PL = 1{,}848.62$ minutes
- $DP = 108.53$ percentage points
- $RP = -47.44$ percentage points

Interpretation of the first five resilience metrics is straightforward. The first, second and third phase lasted 68, 95 and 60 minutes, respectively. The maximum difference on the vertical axis measures 12.80 pp, and the area enclosed by target performance and the resilience curve measures approximately 1,848 minutes. The positive degradation profile of 108.53 pp indicates a convex deviation from a linear degradation, which means performance dropped rapidly due to the cancellation of trains early in the disruption. The negative recovery profile of 47.44 pp indicates a smaller, concave deviation from a linear recovery, which means performance recovered rapidly as many trains could be reinserted shortly after the restart was initiated.

In addition to the obtained resilience curves, performance was calculated individually for each timetable point in the impact area. Minimum performance during the disruption was plotted in the network diagram presented in Figure 5.3. It is observed that performance was affected mostly in the first impact area, where it reached 13.3% locally near the site of the collision. In the second impact area, performance was affected moderately in southwestern direction towards Utrecht (Ut) and in northern direction towards Steenwijk (Swk) and Assen (Asn). This corresponds to the route of the 500/600 series, which is the intercity between Rotterdam and the north of the country. It shows that short-turning these series in Amersfoort and Zwolle had logistical consequences for the rest of the route. In contrast, the impact of the disruption was low in the direction of Hilversum (Hvs) and in eastern direction, where most of the traffic is local and starts or ends in Zwolle. The logistical functionality of Zwolle, which is the second largest hub in the network, proved to be sufficient so that intercity traffic could be rescheduled without interfering with local traffic.

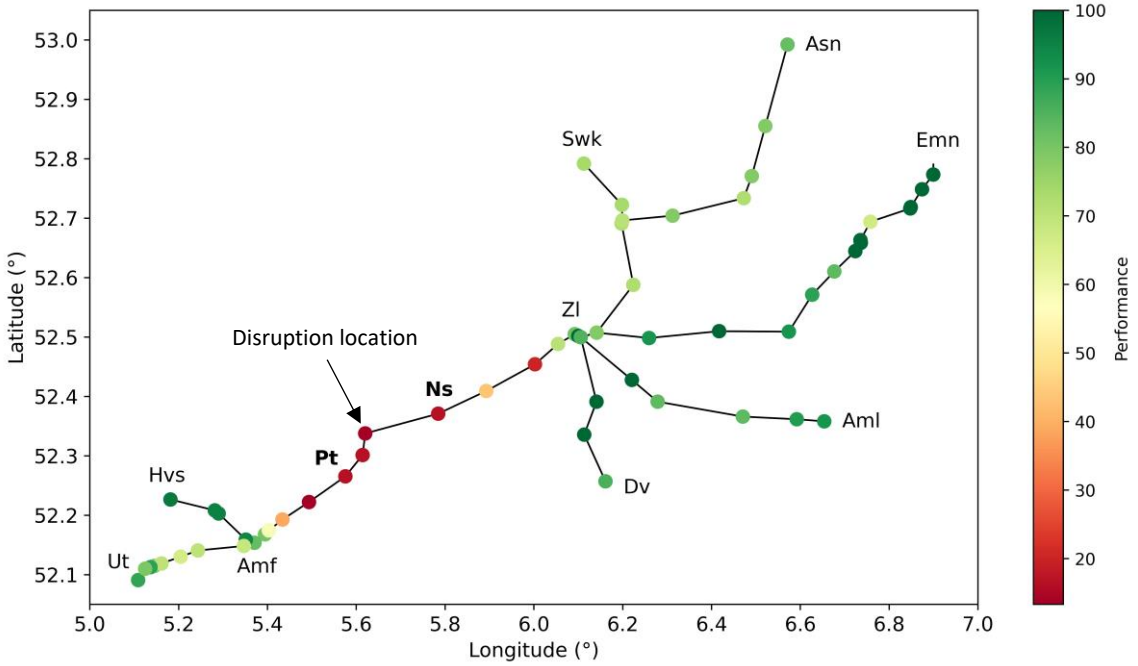


Figure 5.3. Impact area and minimum performance for the studied disruption.

The third impact area was explored as well. This area, which was larger than the first and second impact area combined, appeared to be unaffected in terms of traffic intensity. In terms of punctuality, a drop in performance was observed around $t = 75$ minutes, although no relation to the studied disruption could be identified. Given the size of the area, it may well be the case that the change in punctuality was related to some other, unobserved disturbance.

An even better understanding of the spatial impact is obtained by studying the time-distance diagrams for the first impact area. Diagrams showing the start, middle and end of the disruption are included in Appendix K. The diagrams show that no passenger trains were running between Putten and Nunspeet in the middle of the disruption, while only regional trains were running between Amersfoort and Putten and between Zwolle and Nunspeet. This explains why minimum performance was slightly higher towards the boundaries of the first impact area.

5.3. Representative resilience curves and time differences

After having built trust in the methodology and the chosen performance weight, this section presents the representative resilience curves that were identified in the experiments. This includes the different shapes of resilience curves and the mean and median resilience curve per cause. Differences between the observed and reported timepoints are discussed as well.

Types of resilience curves

The resilience curve is commonly depicted as a bathtub shaped curve with a clearly recognizable first, second and third phase. However, inspection of just a few resilience curves revealed that the actual shape of the curve may not be the same as in theory. Closer inspection of over 100 randomly selected disruptions (at least ten per cause) revealed that it is possible to distinguish between eight types of curves. This number could be slightly lower or higher depending on one's own interpretation. The different types were named as follows:

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

Examples of the types of resilience curves are presented in Figure 5.4. The figures are included in full size in Appendix L. Observations of the sampled disruptions suggest that certain types of resilience curves could be more typical of one disruption cause than of another. The gradual recovery curve appears most typical of collisions, although it was also occasionally observed for train defects. This suggests that collisions may be characterized by a relatively short second phase, although this may be experienced differently in practice considering the required clearing operations. The aftermath curve appears most typical of switch failures, but was occasionally observed for the other infrastructure related causes as well. The late drop in performance in this curve appears to be related to the permanent repair of the infrastructure. Since the area where the mechanics are working must be secured, a higher number of trains may need to be

canceled temporarily, which would cause the drop in performance*. The undefinable curve appears most typical of section/signal failures, which could be explained by the sometimes unclear nature or location of the failure and the fact that this can result in a prognosis which is updated several times*, thereby changing the prospects for the restart repeatedly. This was at least the case for the example shown in Figure 5.4.

The other types of curves do not appear typical of a specific disruption cause, but are still worth discussing. The plateau curve was thought to result from an updated, less restrictive VSM*, but this did not apply to the curves that were observed. However, it may also occur that certain tracks become available again without requiring a new VSM*, which could explain this curve after all. The steady state curve was occasionally observed for particularly long disruptions. An explanation for this type of curve could be that, in certain cases, it is immediately clear to traffic controllers that there will be little traffic for an extended period of time. The timetable influenced curve was occasionally observed for disruptions involving relatively little traffic because they occurred in more isolated parts of the network and/or because they occurred in the early morning or late evening. A longer time window, for example with an interval size of $L = 60$ minutes, would filter out the periodic variation but would also smoothen the curve too much.

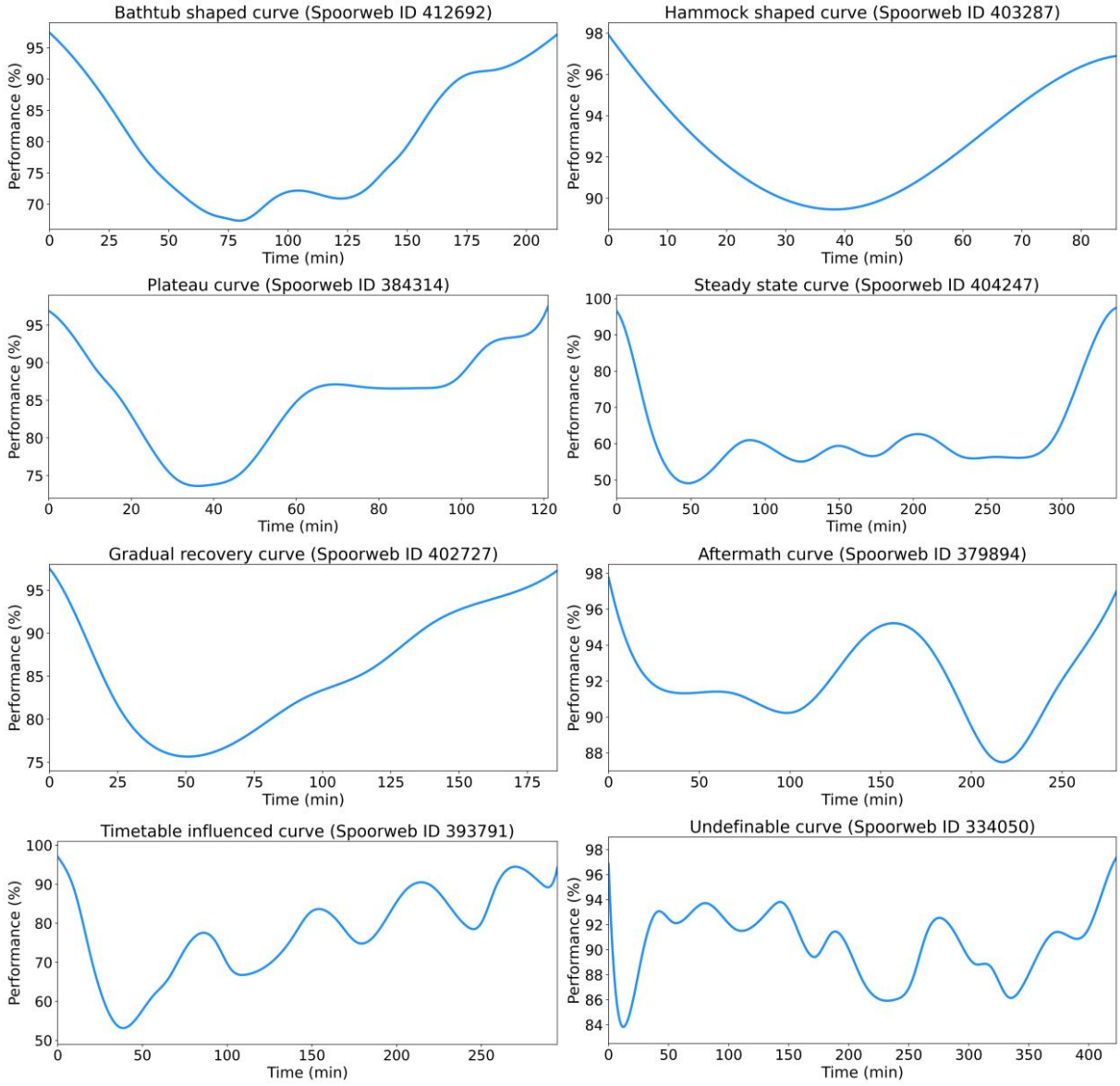


Figure 5.4. Real examples of the different types of resilience curves.

Mean and median resilience curve per disruption cause

To obtain a more general view of the resilience curve, the mean and median resilience curves were drawn for the studied disruption causes. The mean and median performance across a disruption cause were calculated at each time t, where time was expressed as a percentage of the disruption length rather than in minutes. For each curve 101 measurements were taken, from $t = 0\%$ to $t = 100\%$. What this effectively means is that all resilience curves were normalized along the time axis, so they could be presented on the same scale. The mean and median curves are presented in Figure 5.5. The first five plots show the mean and median curve and the central 80% range for each disruption cause for $\lambda = 0.67$. The central 80% range is defined as the range that contains all observations between the 10^{th} and 90^{th} percentile. The sixth plot shows all mean curves in the same plot. The last two plots show the mean curves for the separate contributions of punctuality ($\lambda = 0$) and traffic intensity ($\lambda = 1$). The plots with the mean, median and central 80% range per disruption cause for punctuality and traffic intensity are included in Appendix M.

The plots in Figure 5.5 suggest that differences exist among disruption causes in terms of punctuality and in terms of traffic intensity. A preliminary conclusion with respect to composite performance as well as traffic intensity would be that train defects are the least impactful disruptions on average, while collisions are the most impactful disruptions on average. The difference between collisions and the other causes is most clear in terms of traffic intensity, which is explained by the fact that all nearby trains are typically canceled shortly after a collision is reported. This causes traffic intensity to drop rapidly, which could work through the rest of the disruption as trains are not simply reinserted afterwards. Furthermore, the mean curves for section/signal failures and overhead line failures are quite similar, where switch failures seem to be slightly less impactful on average due to fewer cancellations, and thus, a higher traffic intensity. Differences are also observed in terms of punctuality, which is lowest on average for section/signal failures in the beginning of a disruption. This is explained by the fact that train drivers are often asked by traffic control to reduce their speed and try to “drive the failure out of the system”, which could work if for example some gravel were to be stuck in the insulated rail joint*. The reduced speed automatically causes a decrease in punctuality. The resulting delays can quickly escalate, particularly on busy routes*. Besides being able to show differences between disruption causes, the mean curves for punctuality also disprove the assumption that punctuality is high in the second phase. Punctuality may be high in certain cases, especially in the first impact area when few to no trains are running, but it will be low in other cases as well.

With regard to the shape of the resilience curves in Figure 5.5, it is observed that they do not necessarily resemble the shape of a bathtub. Instead, the curves bear a stronger resemblance to a hammock that is skewed to the left in varying degrees, although the 80% range shows that an arbitrary resilience curve could deviate significantly from the mean curve. The width of the 80% range is relatively small for train defects compared to the other disruption causes. This suggests that train defects are the most consistent type of disruption, which may be partly explained by the fact that as of 2019, extra effort is put by the A&E department into monitoring and evaluating the handling of stranded trains*. Also, train defects result in partial line blockages more often than the other disruption causes. Partial line blockages are generally believed to be less disruptive than full line blockages since a reduced amount of traffic is still possible in those cases. This assumption was supported by a pilot investigation into the mean resilience curve per impact type. With regard to the infrastructure related causes, it is observed that the width of the 80% range is largest in the transition from first to second phase or early in the second phase. This suggests that infrastructure related disruptions are more heterogeneous in terms of degradation behavior than in terms of recovery behavior. For collisions, the width of the 80% range is fairly constant throughout the second phase, which may be explained by the stable conditions during clearing operations.

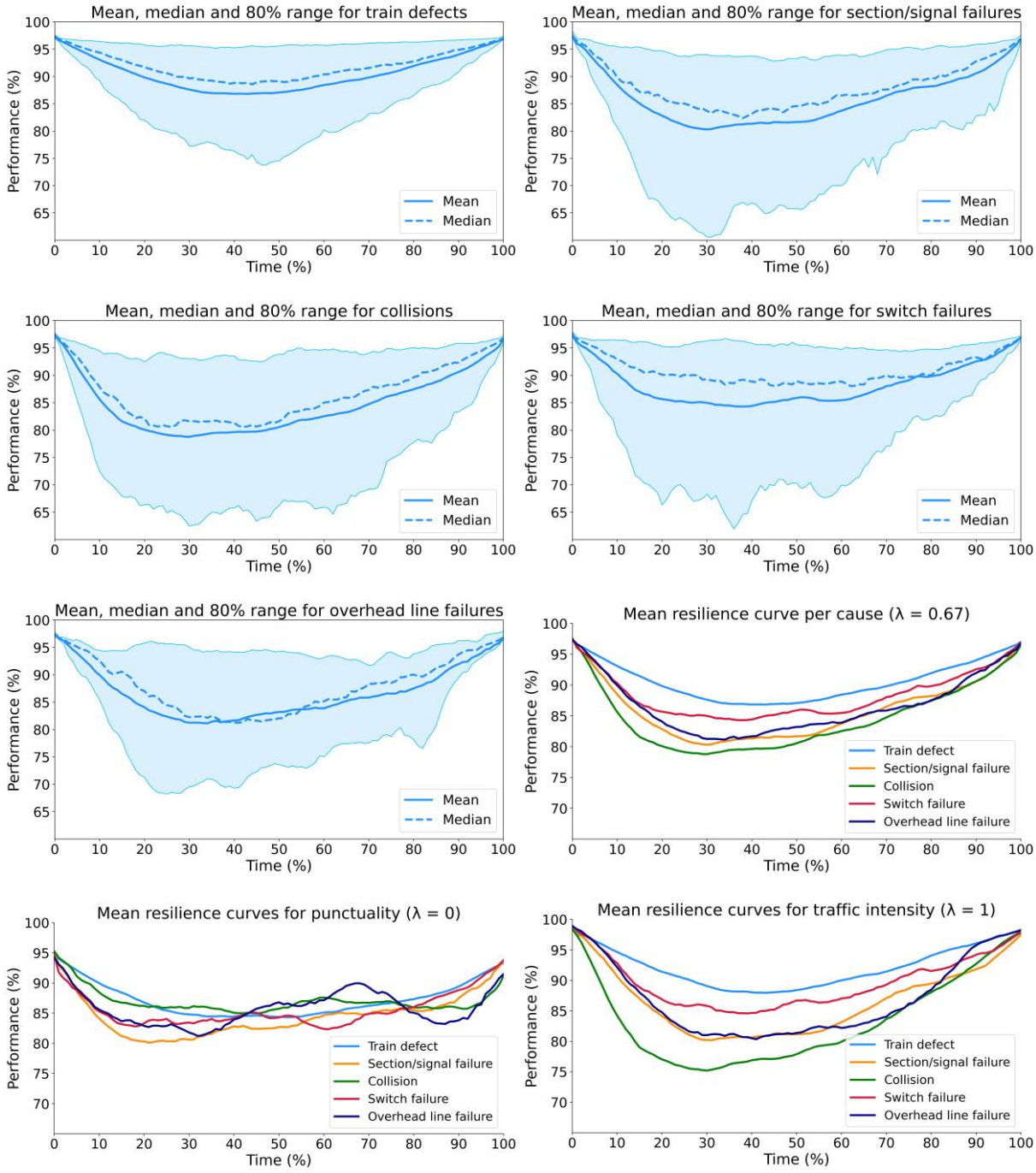


Figure 5.5. Mean and median resilience curve per disruption cause.

Differences between observed and reported timepoints

In addition to the representative curves, the observed and reported timepoints and the total disruption duration were compared among disruption causes. Differences in the timepoints were calculated by subtracting the reported value from the observed value, so: $\Delta T_i = T_i - T_{i,S}$ for $i \in \{0,1,2,3\}$. Thus, a negative outcome indicates that a timepoint was observed earlier than reported, and vice versa. Differences were obtained for 438 of the 445 studied disruptions, as in seven cases the timepoint $T_{1,S}$ was not available. The results are presented in Table 5.3 and Table 5.4 for $\lambda = 0.67$. Table 5.3 presents the mean total duration of the studied disruptions for each disruption cause. Table 5.4 presents an overview of the differences in timepoints. The

tables specify the mean, standard deviation (SD), standard error[5] (SE) and confidence interval (CI) per cause. The standard deviation is a descriptive statistic that measures the variability *within* the sample. In other words, it measures how much the observed values differ from the sample mean. The standard error is an inferential statistic that estimates the variability *across* samples. In other words, it estimates how closely the sample mean is distributed around the unknown population mean. The 95% confidence interval is defined as the mean ± 1.96 SE.

With respect to Table 5.3, it is observed that the studied disruptions lasted approximately 39 minutes longer on average than reported. The differences are largest for train defects and collisions. Collisions were also identified as the longest type of single disruption on average, with a mean observed duration of approximately 261 minutes. In practice though, overhead line failures are known to be the longest type of disruption as they can sometimes span several days. However, many overhead line failures were filtered out in data preparation, for example because of connections to other disruptions or because of the day-night transition that could not be handled in the performance calculation. Thus, the studied overhead line failures are mainly the shorter ones, and the same holds for the other disruption causes. After all, the longer a disruption lasts, the greater the chance that it becomes connected to another disruption.

With respect to Table 5.4, it is observed that differences exist in all timepoints. The mean differences in T_0 are mostly in the order of -10 minutes, which is explained as follows. For train defects, it is known that a train driver will usually try to resolve the cause of the problem by themself before reporting the train defect to traffic control. If the train systems are being reset, the train driver may be unreachable for traffic control in the meantime* while delays start to build up. For collisions, the difference may result from the 30-minute time window used in calculating the resilience curve. Given how rapidly traffic intensity can decrease for collisions, a drop in performance could be observed slightly earlier than it actually occurred. However, it is also known that it may take a while for the dispatcher to inform the control room of a collision, because the dispatcher may give priority to taking emergency measures*. For section/signal and overhead line failures, the differences are likely the result of delays that arise before the disruption is reported. The mean differences in T_1 have positive values, which indicates that the second phase generally does not start directly after the VSM is applied. The difference is smallest for train defects and largest for switch failures. The mean differences in T_2 have both positive and negative values. The difference is in the order of 15 minutes for train defects and overhead line failures, but measures approximately -10 minutes for collisions. The mean differences in T_3 are mostly in the order of 20 to 30 minutes, which indicates that the third phase generally lasts longer than reported. Note that for all timepoints, but especially for T_2 and T_3, the standard deviations are quite large, which indicates that large negative and positives deviations from the mean difference were observed as well. This is partly due to the fact that the steady state detection algorithm may give inaccurate results for unconventional types of curves, but for the most part, the time differences seem to relate to the fact that the reported timepoints are not always logged in time or do not reflect the real state of the system.

Since the reported timepoints do not account for delays, but are based on whether or not trains are running, it could be argued that the time differences are caused by the composite performance indicator. Therefore, the differences were also checked for $\lambda = 1$. In this case, disruptions were found to last only 25 minutes longer on average, which is mainly caused by deviations in T_3. The only other notable deviations (in the order of 10 minutes) were identified in T_0 for section/signal failures and switch failures; in T_1 for overhead line failures; and in T_2 for train defects and overhead line failures. In other cases, the time differences were fairly similar.

[5] To preserve space, the standard error was not included in Table 5.4. However, the interested reader may calculate the standard error based on the information in the table.

Table 5.3. Observed and reported mean disruption duration for $\lambda = 0.67$.

		Observed duration (minutes)					Reported duration (minutes)				
Disruption cause	N	Mean	SD	SE	95% CI		Mean	SD	SE	95% CI	
Train defect	196	153.31	81.85	5.85	141.78	164.84	109.90	70.41	5.03	99.98	119.81
Section/signal failure	97	223.25	99.59	10.11	203.18	243.32	188.10	94.48	9.59	169.06	207.14
Collision	95	261.29	79.53	8.16	245.09	277.50	217.38	72.20	7.41	202.68	232.09
Switch failure	34	238.91	131.51	22.55	193.03	284.80	213.05	126.83	21.75	168.80	257.31
Overhead line failure	16	216.50	54.20	13.55	187.62	245.38	203.41	111.59	27.90	143.94	262.87
Average	438	201.17	99.88	4.77	191.79	210.55	161.95	96.11	4.59	152.93	170.98

Table 5.4. Mean differences between the observed and reported timepoints for $\lambda = 0.67$.

		Difference in T_0 (minutes)				Difference in T_1 (minutes)				Difference in T_2 (minutes)				Difference in T_3 (minutes)			
Disruption cause	N	Mean	SD	95% CI		Mean	SD	95% CI		Mean	SD	95% CI		Mean	SD	95% CI	
Train defect	196	-13.83	22.28	-16.97	-10.69	3.86	22.39	0.71	7.01	14.23	40.91	8.47	19.99	29.58	54.01	21.98	37.19
Section/signal failure	97	-9.72	24.87	-14.73	-4.71	10.88	40.70	2.67	19.08	-1.52	48.63	-11.32	8.28	25.42	44.46	16.46	34.38
Collision	95	-12.89	16.08	-16.17	-9.62	19.93	52.68	9.20	30.66	-9.64	47.34	-19.29	0.00	31.02	51.29	20.57	41.46
Switch failure	34	-2.91	29.54	-13.22	7.40	25.04	58.44	4.65	45.43	7.03	51.15	-10.82	24.87	22.95	84.34	-6.48	52.37
Overhead line failure	16	-11.38	31.77	-28.31	5.56	17.10	32.45	-0.20	34.39	15.97	43.20	-7.05	38.99	1.72	81.05	-41.47	44.91
Average	438	-11.78	22.86	-13.93	-9.63	11.03	39.19	7.35	14.71	5.07	45.93	0.76	9.38	27.44	55.64	22.22	32.67

Table 5.5. Descriptive statistics of the resilience metrics for $\lambda = 0.67$.

		DT (minutes)			RST (minutes)			RCT (minutes)			MI (percentage points)		
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 (percentage points)			RP (percentage points)		
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

Based on the mean resilience curves and time differences, using the composite performance indicator that includes punctuality is believed to be justified. In addition to the arguments provided in Chapter 4, it was found that performance degrades faster on average in terms of punctuality for the infrastructure related causes than for train defects and collisions. Also, section/signal failures and switch failures were found to occur approximately 10 minutes later on average when considering traffic intensity alone than when using the composite indicator. In fact, the standard deviation of ΔT_0 for switch failures is more than double the size for $\lambda = 1$ compared to $\lambda = 0.67$. This indicates there must have been a number of switch failures that were hardly affected in terms of traffic intensity, with a large deviation from the reported start time as a result. Also, it became clear that disruptions last longer on average when punctuality is included, which indicates that delays do occasionally arise in the first and third phase. If only traffic intensity were to be studied, the decrease in punctuality would be overlooked, which is why it is worthwhile to incorporate punctuality in the resilience curve as well.

5.4. Comparison of resilience metrics across disruption causes

This section discusses the group comparisons that were performed with respect to the resilience metrics, where each group represents the disruptions that match one of the five studied disruption causes. First, descriptive statistics of the resilience metrics are reported per group and the assumptions for group comparisons are discussed. Next, the results of the group comparisons are presented for $\lambda = 0.67$ and verified for other values of λ.

Descriptive statistics

To get an overview of the values of the resilience metrics for each disruption cause, descriptive statistics were reported. The mean, median and standard deviation of each metric is presented per cause in Table 5.5. The table shows that a single disruption on average has a degradation time of 49.55 minutes; a response time of 79.17 minutes; a recovery time of 70.61 minutes; a maximum impact of 18.09 percentage points; a performance loss of 2,096 minutes; a degradation profile of -19.54 percentage points; and a recovery profile of -54.46 percentage points. The standard deviations of the metrics are relatively large compared to the means. This suggests that the disruption dynamics may be quite heterogeneous, which is also reflected in the various types of resilience curves that were identified in Section 5.3. Since the first five metrics cannot have negative values, their distributions must be skewed to the left, which means that much larger values than the mean were occasionally observed. In contrast with the first five metrics, the degradation profile and recovery profile are allowed to have negative values as well. Given the size of the standard deviations for these metrics, their mean and median are relatively close to zero. This 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.

Assumptions check

Next, the parametric assumptions for group comparisons were checked. The normality of the dependent variable was checked by performing the Shapiro-Wilk test, which is one of the more powerful normality tests (Yap & Sim, 2011). It evaluates the null hypothesis that the data were drawn from a normal distribution. Test results consist of the Shapiro-Wilk W-statistic and the corresponding p-value. The greater the W-statistic, the more likely it is that the data are normally distributed. The null hypothesis was rejected at a significance level $\alpha = 0.05$ for all metrics across most of the groups. Thus, as an alternative to ANOVA, the Kruskal-Wallis test was considered. Kruskal-Wallis requires independence of observations and equal variances among groups. The first had already been ensured by distinguishing between single and connected disruptions. The second was checked by drawing a kernel density plot for each metric and performing the Levene test, which evaluates the null hypothesis that the data have equal variances among groups. Test results consist of the Levene W-statistic and the corresponding p-value. The greater the W-statistic, the less likely it is that the groups have equal

variances. The null hypothesis was rejected at $\alpha = 0.05$ for all metrics. As a result, the Kruskal-Wallis test could not be used and Welch's ANOVA was selected for the group comparisons. For completeness, the test results and kernel density plots are included in Appendix N.

Results for λ = 0.67

As group comparisons are by definition two-sided tests, the null hypothesis H_0 and alternative hypothesis H_1 for the comparison of resilience metrics among groups are as follows.

H_0: The mean of the resilience metric is the same for each disruption cause.
H_1: The mean of the resilience metric is different per disruption cause.

The results of Welch's ANOVA are presented in Table 5.6, which provides the F-statistic, p-value, η-squared, and whether or not the null hypothesis was rejected[6] at $\alpha = 0.05$. 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. Table 5.6 shows that the first five resilience metrics are significantly different per disruption cause, and that the effect sizes are medium to large. The largest effect size, η-squared = 0.169, was obtained with regard to the performance loss.

Table 5.6. Welch's ANOVA test results for $\lambda = 0.67$.

Resilience metric	F-statistic	p-value	η-squared	H_0 rejected ($\alpha = 0.05$)
DT	4.770	1.77E-03	0.043	Yes
RST	22.352	1.06E-12	0.125	Yes
RCT	9.824	1.45E-06	0.081	Yes
MI	15.954	1.53E-09	0.129	Yes
PL	21.533	6.99E-12	0.169	Yes
DP	0.437	7.82E-01	0.012	No
RP	0.611	6.56E-01	0.008	No

In addition to Welch's ANOVA, the Games-Howell post hoc test was performed. The full results are included in Appendix O. A subset of the results is presented in Table 5.7, which provides the groups A and B, the difference in their means, the standard error, 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 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). It is generally preferred not to use rules of thumb and instead compare the effect sizes to earlier results in similar research (Lakens, 2013). However, such results were 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. For example: based on the last row in Table 5.7, it could be stated that a section/signal failure has a 1463.56 minutes greater per-

[6] The p-value describes the probability that the test statistic would have been as least as large as observed if only chance was at play. 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). Thus, in saying that the null hypothesis is rejected, it is assumed that all other assumptions in the statistical model are correct.

formance loss on average than a train defect, which equals 0.718 times the standard deviation of the observed difference (Hedges' g = 0.718). The probability that a random section/signal failure has a greater performance loss than a random train defect equals 69.5% (CLES = 0.695).

Table 5.7. 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

With respect to Table 5.7, it is observed that train defects have a significantly shorter response time than the other causes, and that collisions have a significantly longer recovery time than train defects and overhead line failures. Train defects have a significantly smaller maximum impact than collisions and section/signal failures. In terms of performance loss, collisions perform significantly worse than switch failures and train defects, and section/signal failures also perform significantly worse than train defects. Medium-sized differences regarding overhead line failures were not found to be significant, which may be explained by the small group size. To conclude, the test results are consistent with the mean resilience curves presented in Section 5.3, which already suggested that collisions are the most impactful single disruptions overall and train defects are the least impactful single disruptions overall.

Results for other values of λ

The results of the group comparisons were verified for other values of λ which are easy to communicate, namely $\lambda = 0.5$, $\lambda = 0.75$, $\lambda = 0.8$ and $\lambda = 1$ (placing equal weight, three times the weight, four times the weight and all weight on traffic intensity, respectively). On average, only marginal changes in the mean values of the resilience metrics were observed. Per group though, some notable changes in the metrics were observed. As λ increases, performance loss steadily increases for collisions, while it steadily decreases for switch failures. The maximum impact of collisions also increases with increasing λ. This underlines the disruptive effect of collisions in terms of cancellations. For $\lambda = 1$, the degradation time and recovery time are shorter on average than for the other values of λ, which is consistent with the shorter total duration if only traffic intensity is considered.

The results of the group comparisons for the selected values of λ are summarized in Table 5.8, which presents the effect size η-squared for each λ and for each resilience metric. Based on this table, a trend may be observed for some of the metrics. As λ increases, η-squared steadily in-

creases for the response time, maximum impact and performance loss. Thus, the more weight is placed on traffic intensity, the more obvious the differences are between groups in terms of these metrics. For $\lambda = 1$, η-squared is considerably lower for the degradation time and recovery time. This means that if punctuality is not considered, the duration of the first and third phase varies less among groups. In fact, for $\lambda = 1$ the null hypothesis could no longer be rejected at $\alpha = 0.05$ with regard to the degradation time. Furthermore, note how there are some inconsistencies in the increasing or decreasing trend in η-squared, for example with regard to the degradation time, maximum impact and performance loss for $\lambda = 0.8$. The explanation for these inconsistencies is twofold. First, it is probable that there are a number of resilience curves which are sensitive to changes in λ because of strong fluctuations in punctuality and/or traffic intensity. As a result, the timepoints may change significantly for slightly different values of λ. Thus, $\lambda = 0.8$ could be a particularly unlucky parameter value for a small subset of disruptions. Second, there are slight differences in the disruptions that were evaluated for each value of λ, since the sample depends on whether a curve stays above target performance; whether a start and end point could be identified; whether a disruption was connected based on the observed timepoints; and whether a steady state could be identified. Even though target performance was adjusted for λ, part of the disruptions did not appear in all samples.

Table 5.8. η-squared per resilience metric for different values of the performance weight.

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

5.5. Relationships between resilience metrics

This section discusses the relationships that were evaluated between the resilience metrics and other explanatory variables for the studied disruptions. First, the results of the regression analyses are presented for $\lambda = 0.67$ and verified for other values of λ. Next, relationships that would have been evaluated but were excluded because of limitations in the data are commented on.

Regression hypotheses

In regression analysis, it is evaluated whether there is a relationship between the dependent variable (DV) and independent variable (IV). This is true when at least one of the coefficients is significantly different from zero. Thus, the hypotheses for each coefficient are as follows.

H_0: The regression coefficient β_i is different from zero.
H_1: The regression coefficient β_i is equal to zero.

Since the normality assumption of the dependent variable was violated for most of the metrics in the group comparisons, it was assumed that the normality assumption of the residual errors was violated as well. Besides, no further effort was made to remove outliers from the remaining set of 445 single disruptions. Therefore, robust Huber regression was selected for regression analysis instead of the standard linear regression. As explained briefly in Chapter 4, a Huber regression model assigns less weight to outliers in the data. The number of observations that are classified as outliers is controlled by the tuning parameter t_h. Large values of t_h decrease the robustness of the regression model against outliers, making it similar to a linear regression

model, whereas small values of t_h increase the robustness of the model against outliers. The default parameter setting in Statsmodels (the Python module used to perform the regression analyses) is $t_h = 1.345$, based on Huber (1981).

Additional data were collected to be evaluated against the resilience metrics. This concerns the customer hindrance per disruption in total minutes delay and the number of train series involved in the third phase. Several candidate DV-IV pairs were selected given the relationships that could logically be expected, such as the response time vs. degradation time, recovery profile vs. response time and performance loss vs. maximum impact. However, based on the scatter plots that were drawn for the regression candidates, there was only reason to assume a relationship between performance loss on the one hand and the maximum impact and total duration on the other hand. This relationship is presented in linear form in Equation (14).

$$PL = \beta_1 MI + \beta_2 TT \tag{14}$$

Where:

- PL is the performance loss
- MI is the maximum impact
- TT is the total duration ($TT = DT + RST + RCT$)
- β_1 is the slope of the regression line with regard to MI
- β_2 is the slope of the regression line with regard to TT

Since performance loss represents the cumulative loss of performance over the course of a disruption, it could be expected to increase with a greater impact MI and longer duration TT. Also, when both the impact and the duration are zero, performance loss should be zero as well, which is why no constant β_0 was included in Equation (14). The relationship of the dependent variable PL with the independent variables MI and TT was evaluated first for the independent variables separately. A Huber regression model was fitted with the default value $t_h = 1.345$. Smaller values of t_h further increase the robustness against outliers, which is why other values were attempted as well, including $t_h = 1$ which is the minimum value in the range that Huber (1981) identified as "good choices". However, the differences were marginal. Subsequently, a multivariate Huber regression model was fitted that includes both independent variables.

Results for λ = 0.67

The results of the regression analyses for $\lambda = 0.67$ are presented in Table 5.9, which provides the variable name, coefficient value, standard error, t-value, p-value and 95% confidence interval. The results in Table 5.9 show that the relationship between performance loss on the one hand and the maximum impact and total duration on the other hand is significant at $\alpha = 0.05$. In the first model, $t(\beta_1) = 64.54$ and r-squared $= 0.677$. The "goodness of fit" r-squared explains the proportion of the variation in the dependent variable that can be explained by the independent variable[7]. In the second model, $t(\beta_2) = 39.83$ and r-squared $= 0.399$. This means that changes in the maximum impact explain a larger part of the variation in performance loss than changes in the total duration, and thus, it could be stated that it is more important to reduce the maximum impact than to reduce the duration of a disruption in order to limit performance loss. The regression lines for the obtained coefficient values are shown in Figure 5.6. For comparison, a Huber regression model with $t_h = 1$ and an ordinary least squares (OLS) model were fitted as well. The figure illustrates that mainly very large observations were marked as outliers, since the Huber regression lines are less steep compared to the OLS model.

[7] Note that r-squared should be interpreted with caution for robust models, as its value will always be lower than the value that would be obtained with an ordinary least squares model.

Table 5.9. Huber regression results for the separate and multivariate models for $\lambda = 0.67$ and $t_h = 1.345$.

Model 1							
	Variable	Coefficient	SE	t-value	p-value	95% confidence interval	
	MI	113.60	1.76	64.54	≈ 0	110.15	117.05
Model 2							
	Variable	Coefficient	SE	t-value	p-value	95% confidence interval	
	TT	9.76	0.25	39.83	≈ 0	9.28	10.24
Model 3							
	Variable	Coefficient	SE	t-value	p-value	95% confidence interval	
	MI	88.88	3.11	28.56	≈ 0	82.78	94.98
	TT	2.82	0.31	9.07	≈ 0	2.21	3.43

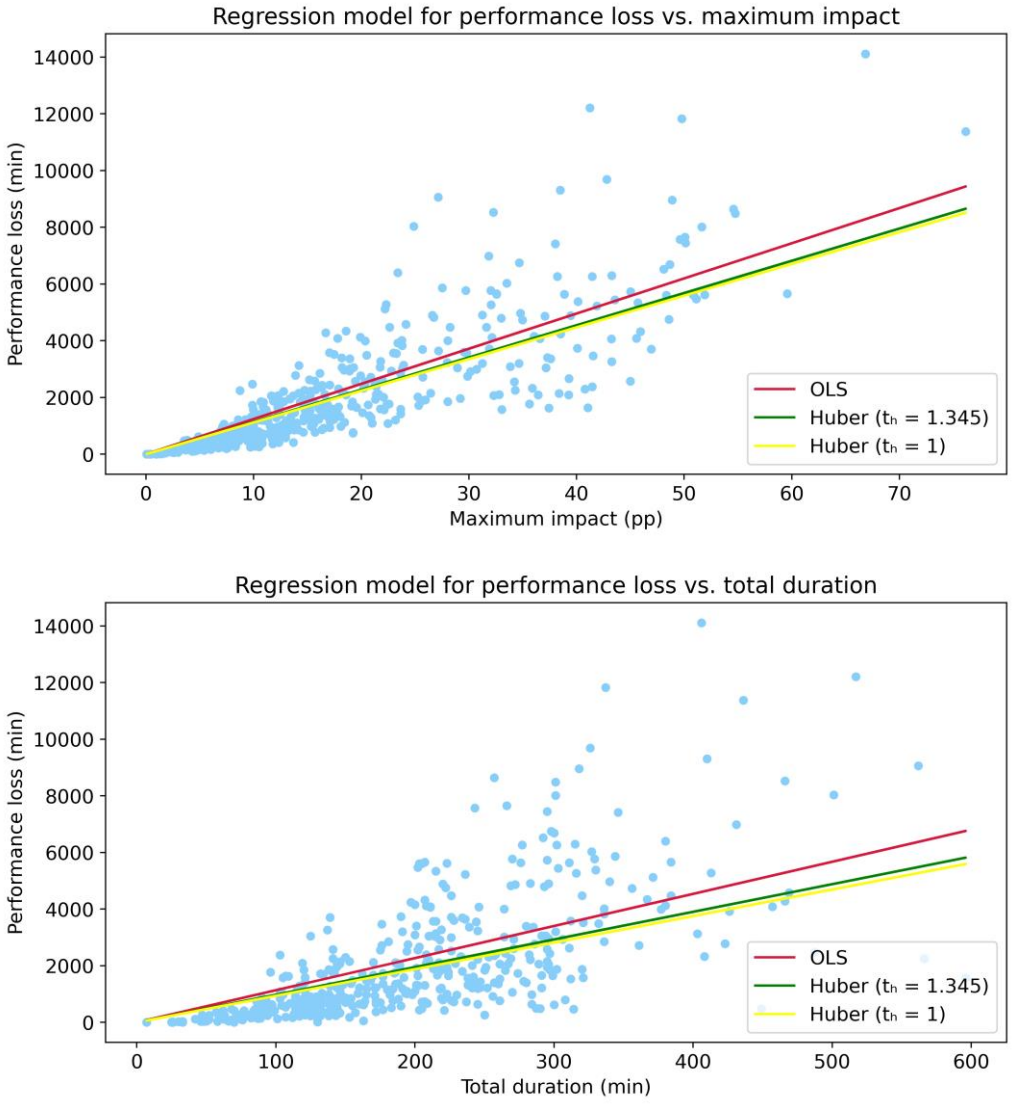


Figure 5.6. Regression lines for a nonrobust OLS and robust Huber regression model.

In the multivariate model r-squared equals 0.709, which means 71% of the variation in performance loss is explained by the maximum impact and total duration. The coefficient values measure $\beta_1 = 88.88$ and $\beta_2 = 2.82$, which means performance loss increases by 88.88 minutes for a one percentage point increase in maximum impact, and it increases by 2.82 minutes for a one minute increase in total duration. Still, 29% of the variation in performance loss is not explained by the multivariate model, which indicates that other factors influence the shape of the resilience curve as well. These could include characteristics of the infrastructure (e.g. the number of switches), timetable (e.g. the train frequency), human action (e.g. the proactiveness of traffic controllers) information supply (e.g. the certainty about the prognosed end time) and external conditions (e.g. the weather). These factors are discussed in more detail in Chapter 6.

Results for other values of λ

The regression results were verified for the same values of λ as used in the group comparisons, namely $\lambda = 0.5$, $\lambda = 0.75$, $\lambda = 0.8$ and $\lambda = 1$. The verification produced comparable results for the other values of λ, though it was observed that r-squared steadily decreases for increasing λ. This occurred in the separate models as well as in the multivariate model. The decrease was mainly caused by a decreasingly strong relationship between performance loss and the total duration. For $\lambda = 1$, around 68% of the variation in performance loss could still be explained by the multivariate model, with $\beta_1 = 90.00$ and $\beta_2 = 2.43$. The fact that β_1 remains relatively stable for increasing λ while β_2 decreases is consistent with the fact that the shorter total duration for larger values of λ is primarily caused by a shorter duration of the first and third phase. These phases typically contribute less to the performance loss since performance is higher than in the second phase, where the maximum impact most often occurs.

Limitations in the data

Additional relationships would have been evaluated if it were not for limitations in the data. The first such relationship is that between recovery time or recovery profile and the number of train series involved in the third phase. It could be expected that a higher number of train series results in a more complex, perhaps longer third phase. The boxplot in Figure 5.7 suggests there may indeed be a positive relationship between the recovery time and the number of train series. This relationship was not evaluated further since it was found that the number of train series in the third phase, as reported in Sherlock, only applies to series that were short-turned and later reinserted. Also, the algorithm in Sherlock only considers the last applied VSM, so if trains were short-turned early in the disruption and this was no longer necessary after updating the VSM, no train series were reportedly involved in the third phase. This explains why 211 (47%) of the 445 studied disruptions were characterized by zero or one train series in the third phase, and why long recovery times were also observed for low numbers of train series.

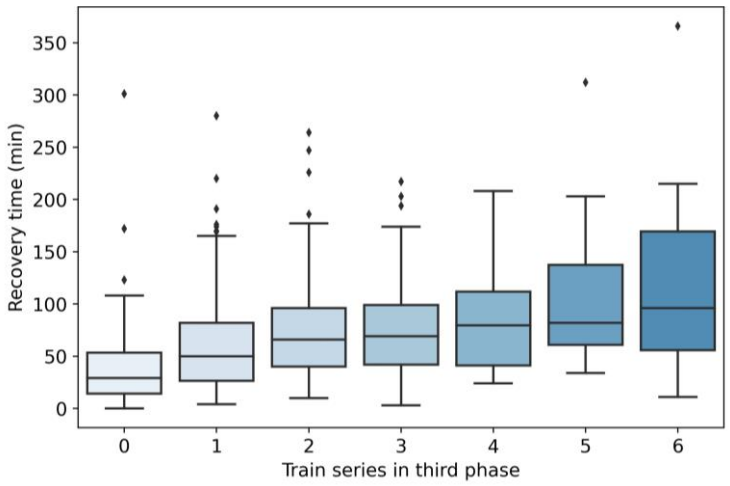


Figure 5.7. Boxplot of the recovery time vs. the number of train series involved in the third phase.

The second relationship that was disregarded is the relationship between customer hindrance and performance loss. Both variables measure the cumulative loss of performance in minutes, although they are calculated differently. Thus, a correlation between the two variables could be expected. Such a correlation would not represent a causal relationship, but would instead be a confirmation that the resilience metric is able to capture the impact of a disruption with reasonable accuracy, assuming that customer hindrance is reported correctly. Yet, based on the scatter plot for these variables, the correlation appears weak. Large values of performance loss were observed for small values of customer hindrance, and to a lesser extent, vice versa. A possible explanation for this discrepancy is that performance loss describes the profile of the resilience curve, and not the total impact on the train service. For example, a resilience curve calculated for a relatively small impact area could have the same performance loss as a resilience curve calculated for a relatively large impact area. In the second case, it is likely that more trains were affected by the disruption, which could lead to more customer hindrance. To overcome this apparent limitation of the resilience metric, the performance loss was normalized for the size of the impact area, but this made hardly any difference. Therefore, the source of the discrepancy should be sought in the reported customer hindrance. Of the 445 disruptions, 53 (12%) were observed with a customer hindrance of zero, but with performance losses up to 4,154 minutes. This deviation can be explained in three ways. First, the allocation of delays to a specific disruption by Monitoring is highly dependent on human input and leaves room for errors and interpretation, especially since input comes from traffic controllers from multiple traffic control areas. Second, it is unknown to what extent the coupling of a Monitoring ID to a Spoorweb ID in Sherlock is correct. Incorrect or missing couplings were observed when a minor disruption (without a VSM) occurred in the same area shortly before the disruption to which the customer hindrance should apply, or when the boundary points in the Monitoring record were different than the ones in the Spoorweb record. Third, delays are only allocated to a disruption until at most 30 minutes after the restart is initiated, whereas performance loss is calculated over the observed duration of a disruption, which includes the entire third phase.

5.6. Resilience curves for connected disruptions

So far, connected disruptions were excluded from the experiments. In an earlier version of the experiments however, the mean resilience curve was drawn for connected disruptions as well, and connected disruptions were also represented in the group comparisons. The results from those experiments led to believe that evaluating connected disruptions as if they were single disruptions could underestimate the impact of a connected disruption. Therefore, this section discusses the resilience curve behavior for connected disruptions separately and explains why connected disruptions should not be compared against single disruptions. First, two connected disruptions of different size are discussed in detail. Next, the factors that can influence the shape of the resilience curve for a connected disruption are summarized.

Table 5.10. Details of the two connected disruptions.

Example	ID	$T_{0,S}$	$T_{3,S}$	Boundary points	Disruption cause	Impact type
1	335628	06:56	13:52	Apd, Bnn	Collision	Full line blockage
	335662	10:33	14:04	Amf	ARI failure	Full timetable point outage
2	359646	12:24	18:28	Asd, Ass	Section failure	Partial line blockage
	359708	15:34	17:37	Sdm, Rtd	Train defect	Full line blockage
	359741	17:22	21:09	Mrn, Ed	Collision	Full line blockage
	359764	18:10	18:33	Asd, Asdl, Ass	Train defect	Partial line blockage
	359775	18:21	19:27	Shl	Train defect	Partial timetable point outage
	359778	18:22	21:21	Ed	Signal failure	Full timetable point outage

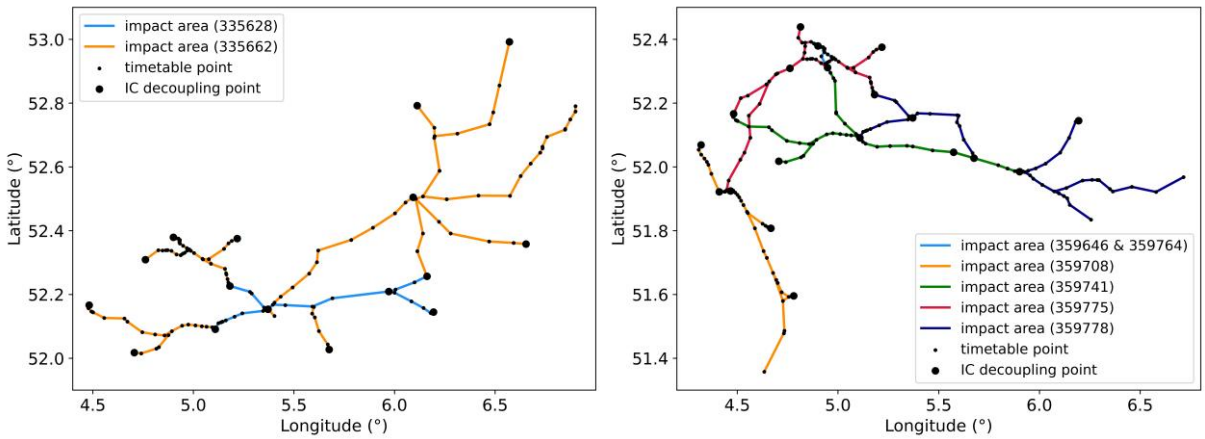


Figure 5.8. Combined impact area of the first (left) and second (right) connected disruption.

The first example concerns a connected disruption consisting of two subsequent disruptions around Amersfoort on December 15, 2018. The Spoorweb ID, reported start and end time, boundary points, disruption cause and impact type of the individual disruptions are presented in Table 5.10. In this example, the first disruption was a collision between Apeldoorn and Barneveld which reportedly occurred at 06:56. Shortly after the Spoorweb record was closed at 10:32 and no further logistical measures were deemed necessary, a failure in the automatic routing system ARI occurred around Amersfoort, which reportedly lasted until 14:04. The combined impact area of this connected disruption contained 146 timetable points, where the impact area of the first disruption (in blue) was a subset of the impact area of the second one (in orange) as shown in the left diagram in Figure 5.8. The resilience curve for the combined impact area is presented in Figure 5.9. The figure shows how the first disruption transitioned into the second one, as performance had already nearly recovered when the second disruption occurred. Due to the aftermath of the first disruption, the second part of the curve starts below target performance, which means no start point could have been identified if the second disruption was studied as a single disruption.

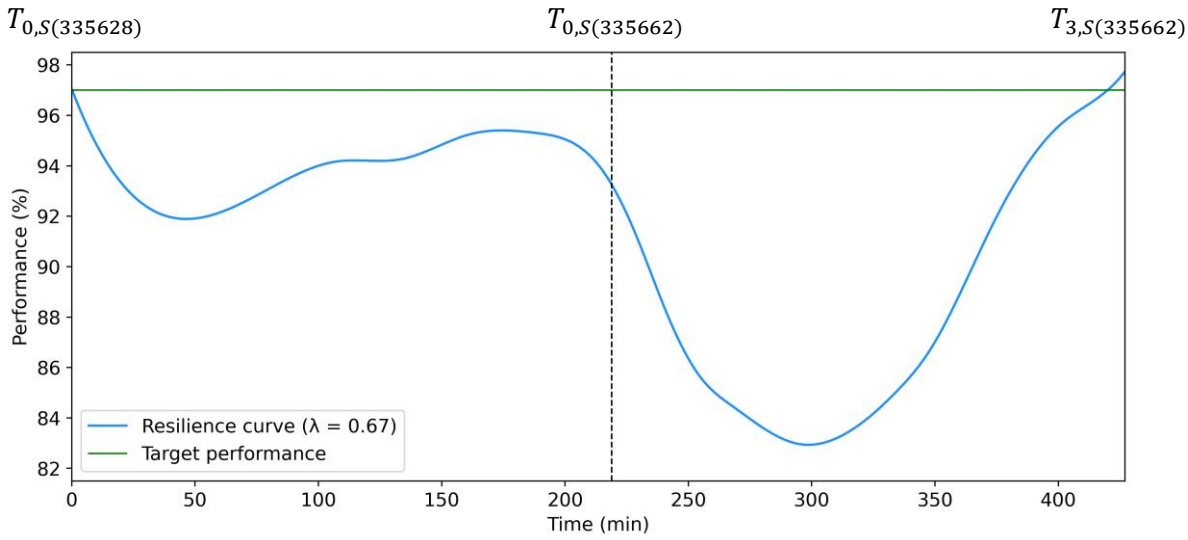


Figure 5.9. Resilience curve for the first connected disruption.

The second example concerns a connected disruption consisting of six disruptions spread across the west and middle of the country on March 11, 2019. Again, the details of the individual disruptions are presented in Table 5.10. The combined impact area of this connected disruption contained 197 timetable points. The maximum area size of the individual disruptions

measured 65 timetable points, so the geographical overlap between part of the disruptions was limited, as shown in the right diagram in Figure 5.8. However, there was significant overlap between the disruptions in Amsterdam (in blue) and Schiphol (in red), which is why those in Amsterdam are barely visible. The resilience curve for the combined impact area is presented in Figure 5.10. Similar to the first example, the first disruption transitioned into the second one. In this case however, the impact of the first disruption was greater than the impact of the second one. The second disruption then transitioned into the third one, which was a collision that caused a full blockage between Maarn and Ede-Wageningen. Three more disruptions occurred in the next hour, including a full timetable point outage in Ede-Wageningen due to a signal failure. As this timetable point was already disrupted due to the collision, the additional impact appears marginal. Around the reported end time of the train defect in Schiphol, the resilience curve starts to recover until eventually the reported end time of the signal failure is reached. Notice that target performance was not yet reached by that time. In fact, target performance was never reached during this connected disruption. Also notice that the impact of individual disruptions can be distinguished in the first part of the curve (until $T_{0,S(359741)}$), whereas this is difficult in the second part of the curve because of the greater overlap in time.

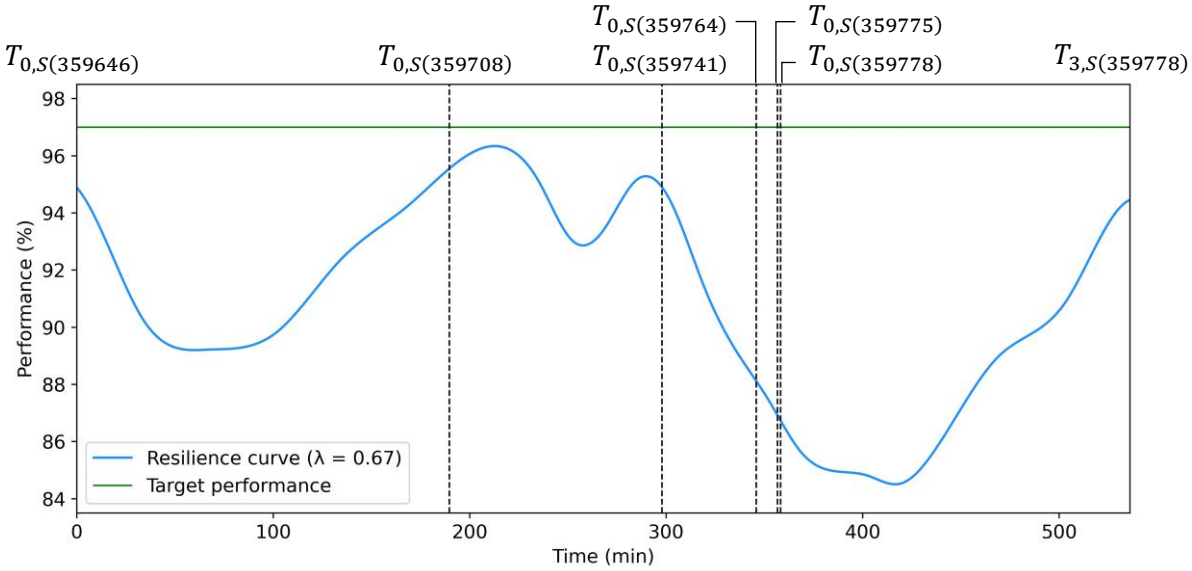


Figure 5.10. Resilience curve for the second connected disruption.

Based on the discussed examples, it could be expected that the resilience curve of a connected disruption can take on many different forms if it is drawn for the total impact area of the individual disruptions combined. For instance, if the first two disruptions in the second example had not occurred, then Figure 5.10 would have resembled the bathtub shaped curve. Thus, the mean resilience curve for connected disruptions (which was found to be similar to the mean resilience curve for train defects) is not thought to be a meaningful nor accurate representation of the disruption dynamics for connected disruptions. Still, studying the resilience curve for connected disruptions or its individual constituents could be useful to learn about the degree in which cascading effects occur and about the performance of traffic control when measures need to account for disruptions elsewhere. Herein, attention should be paid to the following:

- The number of individual disruptions that constitute a connected disruption.
- The extent to which the individual disruptions overlap in time.
- The extent to which the individual disruptions overlap geographically.
- The size of the impact area of individual disruptions relative to each other.
- The geographical separation between the location of individual disruptions.
- The possibility of a causal relationship between two or more individual disruptions.

5.7. Networkwide resilience curves for extreme days

This section presents the resilience curves for the red and black days that were excluded from previous experiments. So far, resilience curves for single disruptions were drawn locally for the affected area. However, on black and near-black days, the influence of simultaneous disruptions or other factors such as the weather could be so strong that this is no longer considered appropriate. Therefore, the curves for extreme days were drawn networkwide for the entire day and were compared to the networkwide resilience curve for a number of nonextreme days in the same period. Although these curves could be inherently different from the curves that were presented earlier in this chapter, they may provide more insight into the potential effects of a disruption outside the first and second impact area.

A distinction was made between extreme days that occur due to a common cause (e.g. snowfall, extreme heat, a strike) or due to an accumulation of disruptions in the network. The reason for this distinction is that the first category can mostly be anticipated in advance, while the second category is of a more coincidental nature, requiring a different approach to disruption management. In particular, it could be argued that traffic control plays a larger role in the second category, since days in this category can hardly be prepared for and require a significant amount of rescheduling during the day. The curves for both categories are presented separately. Resilience curves for extreme days with a common cause are presented in Figure 5.11. Resilience curves for extreme days that occurred due to an accumulation of disruptions are presented in Figure 5.12. As a reference, the range of values was plotted for the networkwide resilience curve on nonextreme days in January, July and November[8].

With respect to Figure 5.11, it is observed that the range of resilience curves for nonextreme days is quite narrow and average performance is quite high. This confirms the assumption that the effects of individual disruptions are typically not well observable on a networkwide scale. This is different for extreme days, which show larger fluctuations in the curve during the day. In addition, the curves for the snowy days in January appear different from the curves for the hot days in July. On January 22 and 30, the timetable was adjusted so that only 80% of the trains were running compared to the regular timetable. Decision making took place one day in advance, which means that the adjusted timetable was not observed in the realization data. As a result, the curves remain relatively stable around 80% in terms of traffic intensity until the start of the evening, when trains appear to have been slowly reinserted in the timetable. The fluctuations in the curves until approximately 19:00 are thus mainly explained by changes in punctuality. The curves for July show a different picture. On July 25 and 26, the timetable was not adjusted, which explains why the curves start relatively high. On both days, no more than two disruptions were reported in the morning, When temperatures began to rise in the afternoon, more disruptions started to occur and the curves steadily degrade. The disruptions in question were almost exclusively train defects and section/signal failures. The curve for March 18 shows a different picture again. The curve starts low because only few trains were running in the morning due to a strike, but for the rest, this was an ordinary day. When comparing the different resilience curves, notice that the curve does not always recover to its original level by the end of the day. This could be explained by the fact that, after a day full of disruptions, it is probable that much of the rolling stock is not in the correct place to resume service the next day or even the same evening. Still, performance at the start of the next day was consistently observed to be higher than the night before, although not necessarily near target performance, which indicates there could be spillover effects to the next day.

[8] The days that were selected as a reference are January 9, July 22 and November 21, 2019, which could be labeled as moderate days in terms of the number of disruptions. Each of these days was followed by two extreme days in the same month.

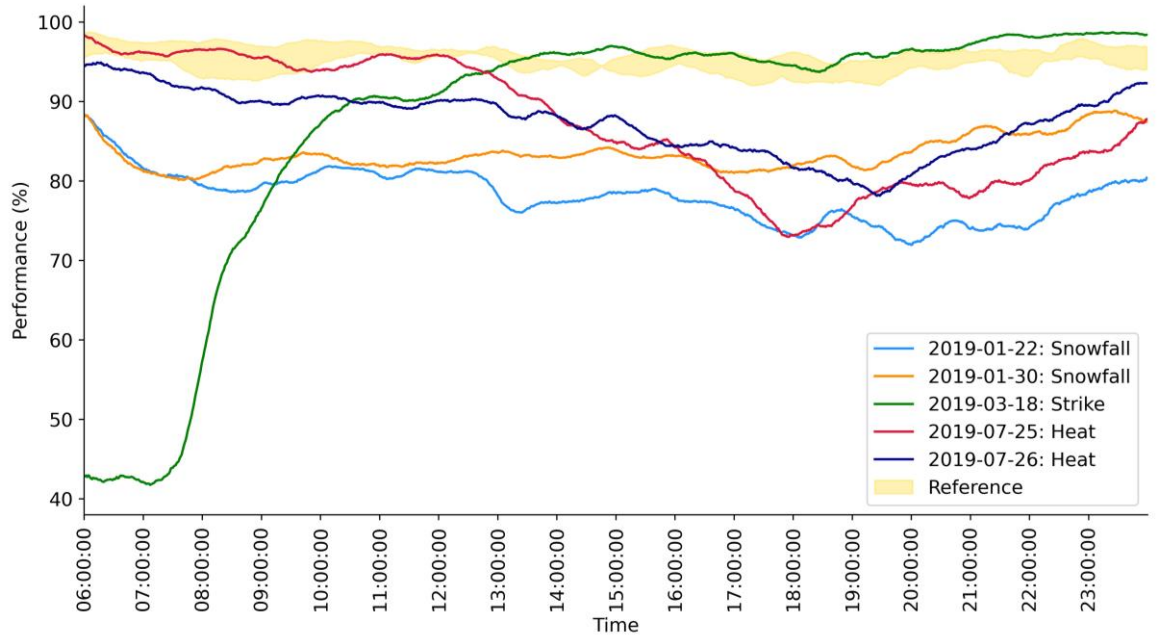


Figure 5.11. Networkwide resilience curves for extreme days due to a common cause.

With respect to Figure 5.12, it is observed that the resilience curves for November 27 and 28 are less exotic than the curve for June 24. It was already highlighted in Chapter 3 that an overhead line failure west of Woerden and a train defect at the station in Woerden occurred shortly after each other on November 27. This caused the blockage of a major artery in the network for a large part of the day. The curve degrades as soon as the overhead line failure was reported at 12:20. It is less clear what caused the fluctuations in the curve for November 28, but a train defect near Amsterdam in the morning and a train defect near Amersfoort in the afternoon are probable explanations. For June 24, the drop in the curve around noon is explained by simultaneous disruptions near Woerden, The Hague and Eindhoven. Then, at 15:18 a collision occurred at the station in Utrecht. While the disruption was being managed, a national failure of the telephone network occurred at 15:42, which made it impossible to contact emergency services and made communication between traffic control centers more difficult. Additionally, two simultaneous train defects were reported around 16:41 on the high-speed line: one north of Rotterdam and one south of Rotterdam. All disruptions combined, this caused networkwide performance to drop as low as 63.5%. This means that at the lowest point in the curve, more than one third of the passenger trains in the entire country was either delayed or canceled.

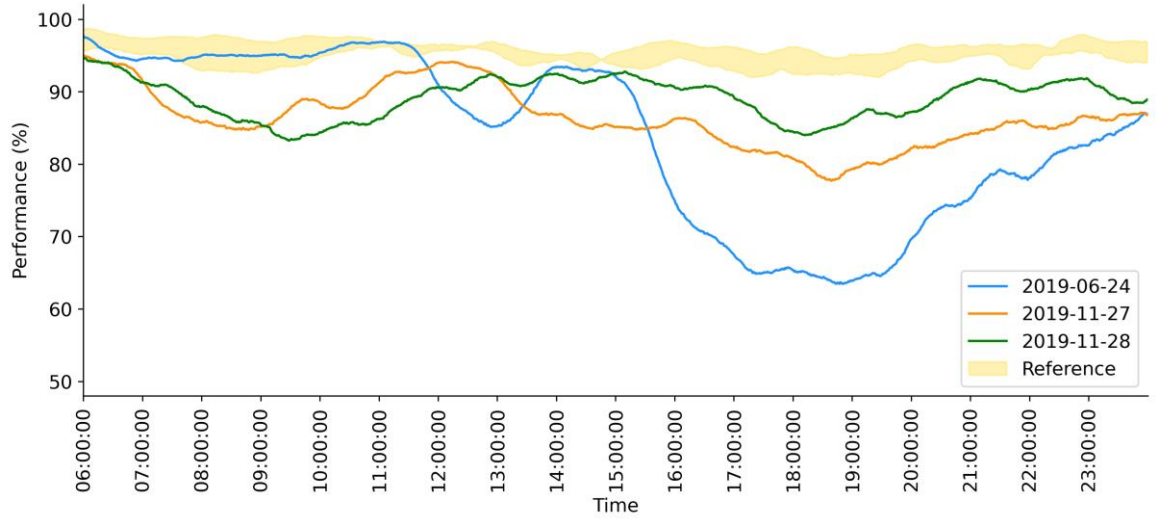


Figure 5.12. Networkwide resilience curves for extreme days due to accumulation.

Based on the resilience curves for extreme days it was concluded that there are multiple ways in which an extreme day could occur. Correspondingly, there are multiple ways in which the resilience curve could behave on those days. The evolution of performance during the day could be explained by studying the curve in terms of the composite performance indicator. The three extreme days that occurred due to an accumulation of disruptions have something in common: they could all be traced back to one or more disruptions in control area Utrecht. With regard to resilience, this raises the question whether or not the railway network is sufficiently resilient against disruptions at the heart of the network, which could be a direction for future research.

5.8. Chapter summary

In this chapter, the resilience evaluation framework was applied on a case study. The results of the experiments in the case study were discussed to incrementally build knowledge about the behavior of the resilience curve, and thus, the dynamics of disruption and recovery in the Dutch railway network. The key findings from the experiments are summarized as follows:

- The observed shape of the resilience curve does not necessarily resemble the shape of the curve as depicted in theory. Instead, eight different types of curves could be identified. The mean resilience curve per disruption cause is shaped like a hammock that is skewed to the left.
- Considerable differences exist between the observed and reported timepoints. Disruptions were frequently observed to start earlier and/or end later than reported. A single disruption was observed to last 39 minutes longer on average than reported when using the composite performance indicator with $\lambda = 0.67$. When only traffic intensity is considered, a disruption was observed to last 25 minutes longer on average than reported.
- The longest resilience phase on average is the response phase, although the recovery phase is not much shorter on average. The mean and median degradation and recovery profile are relatively close to zero, which indicates that the shape of the resilience curve in the transition phases is neither predominantly concave nor predominantly convex on average, but rather linear or mixed.
- Significant differences exist among disruptions of different causes in terms of the duration of the resilience phases, maximum impact and performance loss. Train defects may be considered as the single disruptions with the lowest overall impact, whereas collisions may be considered as the single disruptions with the highest overall impact.
- There is a significant positive relationship between performance loss on the one hand and the maximum impact and total duration on the other hand. The relationship with the maximum impact is the most clear. No additional relationships could be identified between the resilience metrics.
- The shape of the resilience curve for connected disruptions may be quite diverse when studying the combined impact area, which means generalizing connected disruptions does not yield a meaningful nor accurate representation of the disruption dynamics.
- A distinction can be made between two categories of extreme days: those due to a common cause and those due to an accumulation of disruptions in the network. The evolution of system performance during the day can be explained by studying the networkwide resilience curve. The effects of individual disruptions on a regular day are not well observable on a networkwide scale.

Answer to subquestion 4

With the knowledge obtained in this chapter, the fourth subquestion is answered.

Subquestion 4: Which approach should be taken to evaluate railway network resilience for a large and heterogeneous set of disruptions based on traffic realization data?

The resilience evaluation framework proposed in this thesis proved successful for evaluating a large and heterogeneous set of disruptions. The evaluation requires at the very least knowledge of the plan time, realization time and location of a train activity; knowledge of the start and end time, location and cause of a disruption; and knowledge of the structure of the network. The evolution of performance in time can then be calculated as a centered moving average for a specified impact area. A composite performance indicator representing traffic intensity and punctuality may be used to account for the fact that 1) some disruptions perform worse in terms of traffic intensity while others perform worse in terms of punctuality, 2) the two components may develop differently during the course of a disruption, and 3) excluding the punctuality component means that the duration of a disruption could be underestimated. The start and end time of a disruption can be determined based on a performance threshold that represents the average performance on a regular day. Once the performance calculations are completed, the resilience curve can be drawn and inspected visually. The curve may be described quantitatively by a set of resilience metrics, which should represent the multidimensional nature of resilience. Group comparisons may be performed to identify differences between the resilience metrics among groups, provided there are no dependencies among the studied disruptions. This can be ensured by making a distinction between single and connected disruptions. The specific disruption cause may be chosen as the variable that defines group membership. Subsequently, regression analysis may be performed to identify relationships between the metrics, although the presence of such relationships might not be abundant.

6. Discussion

This chapter further elaborates on the results from the experiments, discussing the quantitative but also qualitative aspects of resilience. Understanding these aspects helps create insights into how to prevent, anticipate and mitigate disruption effects. Section 6.1 discusses general remarks based on the experiments and interviews. It describes factors that might influence the shape of the resilience curve and summarizes frequently heard remarks made by respondents. Section 6.2 addresses the practical implications for disruption management practices in the Netherlands. It highlights areas of improvement, discusses the preferred shape of the resilience curve and discusses how the disruption dynamics might change in the near future. Section 6.3 recapitulates the contributions of this thesis to railway resilience theory. Section 6.4 summarizes the chapter and provides the answer to subquestion five.

6.1. General remarks

The general impression from the experiments in Chapter 5 is that the disruption dynamics are quite heterogeneous. Even though differences exist among disruption causes in terms of the resilience metrics, and some types of resilience curves were found to be more characteristic of one disruption cause than of another, there is still significant heterogeneity among disruptions within each cause. Furthermore, a positive relationship was found between performance loss on the one hand and the maximum impact and total duration on the other hand. Since part of the variation in performance loss could not be explained by the regression model, other factors are at play that determine the shape of the resilience curve, and thus, the values of the resilience metrics as well. This includes the maximum impact and total duration which are, after all, inferred from the resilience curve. Based on interviews and expert judgment, such factors may be categorized as characteristics of the infrastructure, timetable, human action, information supply or external conditions. Imagine comparing two disruptions of the same cause, with the same duration and maximum impact, but occurring at different times and/or locations in the network. Factors that explain differences in the shape of the resilience curve between the two hypothetical disruptions might include, but are not necessarily limited to, the following:

Infrastructure

- **Number of railway tracks**. A higher number of tracks improves the network capacity under normal conditions, and logically, under disrupted conditions as well. In case of a partial line blockage, quadruple tracks may still allow bidirectional traffic with one dedicated track per direction, while the partial blockage of a double-track line means that the remaining track has to be used to facilitate traffic in both directions.
- **Number of railway switches**. A higher number of switches enhances the rescheduling options for traffic controllers, as this makes it easier for trains to be stabled or receive a track change. However, it also increases the probability of a switch failure occurring.
- **Network connectivity**. A higher degree of connectivity means that trains are more easily rerouted, although this is uncommon for passenger trains. However, it also means that the impact area may grow more quickly and that disruptions may become connected more easily.

Timetable

- **Number of train series**. A higher number of train series, possibly from different TOCs, means that more trains are heading in different directions, which could increase the spread of disruption effects outside the disrupted area. It might also cause a more complex and slower restart due to the setting of priorities as to which train series go first.
- **Train frequency**. A higher frequency means that secondary delays might arise and propagate more easily because there is less space between trains.

- **Ratio of intercity traffic versus regional traffic**. Regional traffic might experience fewer logistical consequences than intercity traffic, since regional trains run on shorter routes and can be short-turned at more locations in the network. Intercity trains are only allowed to short-turn at stations with stabling capacity.
- **Number and length of freight trains**. More and longer freight trains can make it more difficult to temporarily stable them on a siding so that priority can be given to passenger traffic, at least until the moment of restart.

Human action

- **Experience of actors in the disruption management process**. A more experienced traffic controller, dispatcher, general controller, first responder, mechanic etc. may be more likely to take swift and decisive action.
- **Proactive attitude of traffic controllers**. A more proactive traffic controller may be more likely to take logistical measures early in the disruption in an attempt to prevent secondary delays.
- **Time pressure to reach the second phase**. A higher time pressure to complete all of the preparations for the VSM may result in a suboptimal adjustment to the traffic plan, which may cause performance to be unnecessarily low in the second phase.
- **Time pressure to restart the train service**. A higher time pressure to restart the train service may lead to a restart that is improperly prepared and that may be initiated when not all required rolling stock and/or crew are in place yet.

Information supply

- **Swiftness of reporting the disruption**. The quicker a problem with the rolling stock or infrastructure is reported (regardless of whether it eventually classifies as a disruption), the quicker logistical measures can be taken to prevent secondary delays.
- **Swiftness of communication throughout the chain**. The quicker the details of a disruption are known to each actor in the disruption management process, the quicker each actor has the critical information needed to fulfill their role.
- **Clarity about the cause and location of a disruption**. The more clear the exact cause and location of a disruption are, the more clear it is to traffic controllers and dispatchers which measures to take and whether or not to start working towards a VSM.
- **Availability of a VSM**. When a suitable VSM is available, the actions taken in the first phase may be more consistent than if no suitable VSM were available, since in that case, traffic has to be managed in parallel with preparing an adjusted VSM. Once an adjusted VSM is applied, it could lead to a less steady and suboptimal second phase.
- **Certainty about the prognosed end time**. The more certain the prognosed end time is, the easier it becomes for TOCs to adjust their rolling stock and crew planning while awaiting a feasible restart plan.
- **Knowledge about rolling stock and crew planning**. The more knowledge traffic control has about rescheduling by the TOC, the less likely it is that suboptimal measures are taken or that the restart plan is applied prematurely or proves to be infeasible.

External conditions

- **Time of day**. Depending on the time of day, it may be necessary to update a VSM during the disruption or to postpone the moment of applying or ending the VSM, for example until after rush hour.
- **Weather**. Suboptimal weather such as heat, snowfall or frost might make it more difficult to reach the disruption site, perform repair works, or guarantee the feasibility of the VSM without delays.

Additionally, a general picture of (the issues with) disruption management in the Netherlands was created based on frequently heard remarks from respondents. Most notable was the contradiction between reactive and anticipatory disruption management. Although Schipper and Gerrits (2018) ranked disruption management in the Netherlands as highly anticipatory due to its reliance on contingency plans, the plans only apply to the second resilience phase. In the first phase, disruption management remains highly reactive due to the absence of guidelines, and traffic controllers and dispatchers rely mainly on their experience. It was mentioned that the increase in workload in the first phase in combination with the time pressure is intense. At the same time, the response at the very start of a disruption appears to be rather passive, even though acting more proactively might spread the workload. In preparation of the third phase, it appears to be hard to make an accurate prognosis of the end time, which determines the moment of restart. Uncertainty about the prognosis might result in suboptimal decisions with regard to timetable, rolling stock and crew rescheduling, for example if crew are not available in the right place when the prognosis is advanced. In fact, the cluelessness of ProRail's traffic control about the rolling stock and crew planning from NS' side in general was another notable and recurring remark. This might contribute to the fact that the restart plan frequently does not appear feasible after all, even though the plan is communicated with the TOCs in advance, which in turn could result in a longer third phase.

6.2. Practical implications

The findings from this thesis could have practical implications for disruption management in the Netherlands, which are discussed in this section. First, current areas for improvement are discussed, where the focus is placed on each of the resilience phases and on the handling of collisions in general. In this part, it is also discussed whether a particular shape of the resilience curve should be preferred and how this relates to the application of a VSM. Last, an outlook to the future is given by describing how the disruption dynamics might change as a result of future developments, specifically the shift towards high-frequency railway traffic and the implementation of the new train protection and signaling system ERTMS.

Areas for improvement

First phase

Rescheduling in the first phase is still highly reactive, as there are currently no predefined solutions for this phase. Traffic controllers have to rely on their experience and are faced with a high workload, which leads some to say there is often "no time to think". Taking logistical measures becomes more difficult when a problem is not reported immediately*, information supply is limited*, there is a dependency on the traffic controller of the neighboring traffic control area* or NS' RBC cannot be reached*. From practice it was already known that train defects are often reported late. However, in the experiments it was found that the other types of disruptions (except for switch failures) are also reported late on average. This could be the result of a late notification or because of the decision by traffic control to wait and see how the situation develops. Yet, for every minute that is lost, delays can arise, which may spread to other trains. It is therefore important that trains are canceled if necessary to prevent secondary delays. Herein, it is also important that the notifier communicates the available information quickly throughout the chain of actors involved in the disruption management process, regardless of how relevant this information is to themself. In general, it is believed that there is always at least one actor in the chain who possesses critical information*, but may not realize the value of this information to others.

To reduce the workload for dispatchers and traffic controllers in the first phase, it would be favorable to design predefined solutions for this phase as well. The standardization of measures could help streamline the rescheduling process*, which may also relieve some of the time pressure. On this note, it is important to raise awareness among practitioners about the

fact that getting quickly into the second phase does not necessarily tell much. In fact, the experiments showed no relationship between the degradation time and any other resilience metric. Thus, the purpose of predefined solutions for the first phase should not be to reduce the duration of this phase, which is already the shortest on average, but to allow traffic controllers to come to a well-advised decision about a feasible traffic plan for the second phase.

Second phase
Improvements in the second resilience phase are inextricably related to the design and application of VSMs. Considering the current working practices, the existing VSMs are probably already as detailed as predefined solutions can be. More detailed VSMs that specify the train numbers instead of train series and account for real-time traffic conditions in the network would require a mathematical optimization model that can design tailor-made VSMs. Still, it could be worth investigating to what extent the current set of VSMs can be expanded, provided there is enough capacity in terms of personnel to design and manage the VSMs. Given the high number of connected disruptions that were excluded from the experiments, it might be good to assess if there are locations in the network where connected disruptions are common, and if combinations of disruptions exist which deserve dedicated VSMs covering a larger area.

The fact that VSMs are currently still predefined does not mean that performance develops the same in all cases, as is illustrated by the different types of resilience curves and the large standard deviations in the resilience metrics. This raises multiple questions, such as: should a certain type of resilience curve be preferred? If so, which one is it? And how could it be achieved? The effect of a VSM varies per disruption and depends on how well the VSM is prepared, how closely it is followed and how well the transition towards the restart plan is arranged. Since a VSM does not dictate the actions for each specific train, but merely prescribes the actions per train series, traffic controllers are free to deviate from the VSM during the disruption. However, if predictability towards the TOCs and the passengers is valued as one of the rescheduling objectives, then deviating from the VSM is not preferred. To promote predictability, the bathtub in its essence could still be regarded as the preferred outcome of the rescheduling process, although performance might deviate from the bathtub shape in reality. The hammock shaped curve, the bathtub shaped curve and the steady state curve could be regarded as the preferred types of resilience curve, representing the short, moderate and long bathtub, respectively. In contrast, undesirable types of curves include the gradual recovery curve and the undefinable curve, which appear to be caused in broad terms by too many cancellations in the beginning and too much uncertainty around the prognosed end time, respectively.

Regarding the shape of the resilience curve in the transition phases, the curve should ideally be made as concave as possible. A concave degradation would mean that performance degrades gracefully to an acceptable level in the first phase, where a concave recovery would mean that many trains are reinserted simultaneously and without delays in the third phase. In terms of the resilience metrics, this means that the degradation profile and recovery profile should ideally have as large negative values as possible, although a strongly concave profile might not be realistic. Still, a bathtub with a concave first and third phase could be achieved by taking sufficient, but not overly rigorous first-phase measures as soon as possible; preparing a structurally feasible VSM which is respected throughout the second phase; and preparing a structurally feasible restart plan that allows as many train series as possible to be reinserted with a limited risk of delays.

Third phase
Similar to the first phase, there are currently no predefined solutions for the third phase, aside from the restart framework that is included in a limited number of VSMs. Adjustments to the traffic plan in the third phase are therefore the result of the reactive capacity of traffic controllers. Yet, in the interviews it was often mentioned that traffic control has limited influence on

(and limited interest in) how the third phase develops. On the one hand, it could be argued that a traffic controller is not responsible for the product that the NS or other TOCs provide*. On the other hand, ProRail does have an indirect interest in how the third phase develops, as it is not only responsible for traffic management, but also supervises the utilization of network capacity*. Once in the third phase, it may be hard to make drastic changes to the traffic plan in case the restart does not go as planned. After all, canceling a train is an autonomous decision by the traffic controller, but reversely, the traffic controller cannot simply decide that a train should run again, as this depends on the availability of rolling stock and crew. Thus, improvements in the third phase should focus on the preparation of this phase, specifically regarding the prognosed end time and the restart plan.

Making an accurate prognosis of the moment when the infrastructure will become available again is already a first step towards preparing a feasible restart plan. When there is certainty about the prognosis, which means it is not likely to be postponed or advanced, TOCs can adjust their rolling stock and crew planning accordingly. Advancing the prognosis should not necessarily imply advancing the restart by an equal amount of time, as TOCs may not yet have their rolling stock and crew in place and a feasible restart plan may not yet have been devised. Again, getting quickly into the next phase does not necessarily tell much, especially since the only time pressure comes from the steady increase in customer hindrance. Instead, it should be more important to come to a well-advised decision about a structurally feasible restart plan.

A structurally feasible restart plan could be achieved by better communication between ProRail and the TOCs, specifically NS' RBC. In the current process depicted in Figure 3.5 communication already takes place, but in practice, traffic control has limited knowledge of the availability of rolling stock and crew. At the same time, the NS assumes that changes to the traffic plan are feasible in most cases*, when in fact, the restart plan is frequently not executed or infeasible after all*. This mismatch between knowledge and expectations might cause a longer third phase in which new delays can arise, as was observed in the experiments. Including a restart framework in the VSM more often could help keep the lines of communication short, since in that case, the restart plan has already been prepared in outline. Additionally, traffic control may want to be more in the lead during the third phase to ensure a smooth execution of the restart plan. This is eventually beneficial for all actors, including ProRail, as it may prevent or at least reduce imbalances in the network that could result from an improper restart.

Collisions

The experiments showed that collisions are the most impactful type of single disruption. This finding is intuitive and is consistent with practice. Collisions are also relatively common, being the third most frequently occurring disruption cause where a VSM is applied. Yet, relatively few improvement efforts seem to be made regarding traffic control during collisions, since it is argued that traffic control has limited influence on the course of action after a collision*. This applies particularly to collisions with a person, which involve actors such as emergency services, forensic investigators and funeral providers. These actors also need to have safe access to the collision site*. Since they mainly operate in the second resilience phase, it is argued here that improvements can be gained in the first phase, specifically regarding the swiftness of reporting the collision and the number of trains that are canceled as first-phase measures. Since it is not always clear from the start if the collision is with a person or not, the worst-case scenario is normally assumed, which is why these improvements apply to collisions in general.

With regard to the swiftness of reporting a disruption, it was found that collisions too are reported late on average. This may be caused by a dispatcher who gives priority to taking emergency measures such as deactivating ARI and calling nearby train drivers*, a dispatcher who awaits details provided by the train driver* or a train driver who is in shock*. However, the sooner a disruption is known, the sooner adequate logistical measures can be taken. If the time

pressure and emotional stress are too high in the heat of the moment, it should be considered to allow a quick and easy way of reporting a collision, such as an alarm button in the driver cabin. The dispatcher already has such a button to halt all traffic in the vicinity*.

With regard to the number of canceled trains, it was found that the gradual recovery curve is frequently observed for collisions, including collisions with a person. This may explain why the maximum impact occurs relatively early compared to the other disruption causes as shown in Table 6.1, which presents the mean and median moment of maximum impact per cause in terms of the normalized duration for $\lambda = 0.67$. The gradual recovery curve makes it seem as if there is no second phase, although a clear second phase would normally be expected due to the necessary clearing operations. Hence, the gradual recovery curve may in fact be a distorted version of the bathtub shaped curve, where unnecessarily many trains are canceled in the first phase. It might thus be worthwhile to critically assess the number of trains that are canceled in the first phase of a collision, or potentially, the area in which trains are canceled.

Table 6.1. Mean and median moment of maximum impact per disruption cause.

Disruption cause	Mean moment of max. impact (%)	Median moment of max. impact (%)
Train defect	44.7	43.5
Section/signal failure	39.7	37.0
Collision	38.7	31.5
Switch failure	50.2	49.0
Overhead line failure	47.5	44.0
Average	42.8	41.0

If no further improvements can be gained in reducing the impact of collisions, attention should be paid to ways by which to prevent collisions from occurring in the first place. One solution could be to better monitor persons on or nearby the tracks, for example by increased camera surveillance. Another solution could be to make the tracks less accessible, for example by further reducing the number of level crossings, placing more and/or higher fences next to the tracks and installing platform doors in stations, although the latter may be impractical due to the varying distance between the train doors per train type.

Outlook to the future

High-frequency rail

To meet the growing demand for passenger and freight traffic by rail in the coming decades, the High-Frequency Rail Transport Program (*Programma Hoogfrequent Spoorvervoer, PHS*) was developed by ProRail with the aim to offer one train every ten minutes on the busiest parts of the network. PHS is currently being implemented, and as a result, the disruption dynamics identified in this thesis might change in the future. Before potential changes are discussed, some background information on PHS is provided.

The first frequency increase was implemented in December 2017 on the A2 corridor between Amsterdam and Eindhoven, where three separate train series (each with a frequency of two trains per hour) together offer one intercity train every ten minutes. Until 2028, PHS will be implemented on six more corridors shown in Figure 6.1, including the freight corridor between the Port of Rotterdam and Venlo (shown in orange). The operationalization of PHS varies per corridor. For example, Breda-Tilburg will have four instead of two intercity trains per hour in each direction, whereas Rijswijk-Rotterdam will have eight instead of six intercity trains and six instead of four regional trains per hour in each direction. Preparations for the RoSA corridor between Rotterdam and Arnhem via Schiphol are currently ongoing, Note that this corridor is not shown as such in Figure 6.1.

With the implementation of PHS the railway network will become even busier, which is why additional infrastructure is being built to facilitate the extra trains. The higher frequency in parts of the network means that traffic controllers and dispatchers will have to be more alert, as the shorter time between trains requires a quick response. The fact that the workload in the first phase is already high emphasizes the need for predefined solutions that apply to the first phase. It will also be increasingly important for traffic controllers to be proactive and to cancel a train early on if required to prevent secondary delays, as delays may propagate more easily given the higher number of trains. This way, a traffic controller creates perspective for themself and for the passenger*. After all, canceling a train at the start of a corridor has fewer logistical consequences than canceling a train midway. As a result of the higher frequency, the resilience curve might degrade faster during the first phase, and the initial impact might be greater in case no immediate action is taken. In the third phase, the resilience curve might recover more slowly because reinserting more train series could ask for a more cautious and stepwise restart to prevent delays. After all, if delays were to arise and spread during the restart, some trains might need to be canceled again, which counteracts the purpose of the third phase.

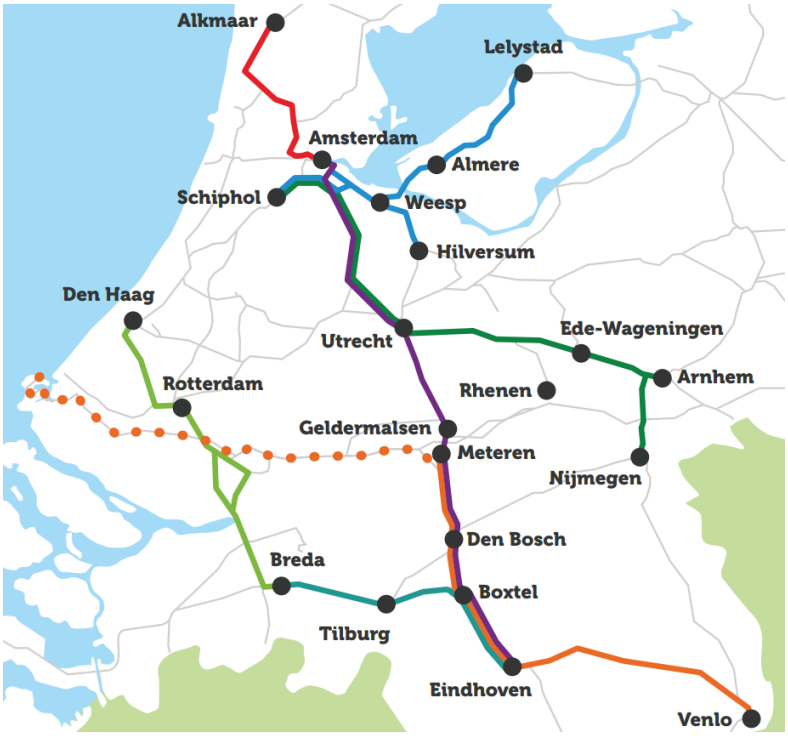


Figure 6.1. Overview of the corridors where PHS will be implemented (ProRail, 2018).

ERTMS

Another development that might change the disruption dynamics identified in this thesis is the implementation of the European Rail Traffic Management System (ERTMS), which is the new automatic train protection and signaling system that has already been partly implemented and continues to be implemented throughout Europe. Some background information on ERTMS is provided before potential changes in the disruption dynamics are discussed.

ERTMS was designed to increase railway traffic safety and standardize train control systems, which should provide better interoperability of passenger and freight railway traffic across the European borders. The main components of ERTMS are the European Train Control System (ETCS) and Global System for Mobile Communications for Railways (GSM-R). In addition, the European Traffic Management Layer (ETML) was defined, but ETML has never been designed in detail nor operationalized. Still, ETML could have potential in the long term (European Rail Research Advisory Council, 2020). ETCS is the part that ensures the safe movement of trains.

It consists of both onboard equipment and trackside equipment. ETCS functionalities include automatic train protection, signaling and positioning. There are three different levels of ETCS which determine the degree of autonomy of the system. ETCS level 2 was chosen to be implemented on the busiest parts of the Dutch railway network in the coming decade. On the high-speed line between Amsterdam and Belgium and on the Betuweroute, ETCS has already been installed. At level 2, trackside signals are no longer required; the train position is determined through beacons in the tracks; and the movement authority (i.e. the permission for a train to move to a specific location, with supervision of speed, and under the constraints of the infrastructure) is transmitted automatically via GSM-R. For a more detailed explanation of ETCS level 2 and the Dutch legacy automatic train protection system, the interested reader may refer to Goverde et al. (2013).

The prospects of ERTMS for traffic control and ICB are just starting to be explored in ProRail. With ERTMS, headway times will decrease and the braking curve becomes more efficient, as trains are able to brake later and accelerate sooner. In case of delays, the infrastructure occupation is expected to improve (i.e. decrease) compared to the legacy train protection system under normal as well as disturbed conditions (Goverde et al., 2013). Assuming that train frequencies remain the same, this means there will be more room for rescheduling. However, as the created space will be likely be filled with more trains, this assumption does not hold. Furthermore, as long as the number and location of railway switches do not change, the logistical handling of a disruption is not expected to change either*. The real-time speed and positioning however will provide traffic controllers with a faster and more accurate source of information, which could be useful for recognizing imminent disruptions and determining the exact location of a defective or stranded train. By acting quickly on this information, delays could be prevented and the run-up at the start of the resilience curve could be reduced. Also, since train locations are known more accurately, the alarm area for a collision could be reduced, which in turn could reduce the initial impact because fewer trains need to be canceled. In the second phase, it could become easier to still have limited traffic in case of a partial blockage, as train drivers can rely on the movement authority that is displayed on their screen instead of worrying about driving on a track that they would normally perceive as “the wrong side”. Eventually, at ETCS level 3, section/signal failures will take on a different nature or perhaps disappear completely. Already at level 2, signals are no longer needed, and at level 3, the same goes for the insulated rail joints and axle counters that currently determine whether or not a section is occupied by a train. This means that ERTMS would not only improve resilience from an operational perspective, but also as a result of the lower complexity of the infrastructure.

6.3. Contributions to resilience theory

Besides having practical implications, the results from this thesis also contribute to the theoretical understanding of resilience, in particular with respect to railways. The contributions are summarized by reflecting on the research gaps that were identified in Chapter 2. The broader application of the results is discussed as well.

Reflection on the research gaps

Research gap 1: The evolution of railway system performance during the consecutive resilience phases is not well understood for disruptions of varying scale and origin.

A general finding from this thesis is that resilience is indeed a multidimensional construct, as many factors were identified that might affect the shape of the resilience curve. It could be stated that resilience is not an intrinsic property related to just one category of factors, but rather, it is an emergent property of the system as a whole. Consequently, the resilience curve shows a range of different behaviors, some of which are more characteristic of one disruption

cause than of another. The different types of curves also show that the bathtub model is an oversimplified representation of the resilience curve, and that in most cases, the curve behaves more dynamically.

In contrast with the existing body of research, the resilience curve was defined in terms of a composite performance indicator that accounts for cancellations as well as delays. This made it possible to capture the interaction between traffic intensity and punctuality, although more research is needed to understand this interaction in detail. The composite indicator proved to be particularly useful to account for delays that arise at the start or end of a disruption, which would be overlooked if only traffic intensity were to be studied. Also, it was shown that the start and end of a disruption can be identified when target performance for near-normal conditions is known, although target performance may not be achievable in all cases.

To describe the profile of the resilience curve quantitatively, two new resilience metrics named the degradation profile and recovery profile were introduced which had not been applied in a railway context before. Given the mean and median value of these metrics, it was concluded that the resilience curve is neither strongly concave nor strongly convex on average. Other than that, the added value of the metrics was limited since no significant differences among groups were found in terms of these metrics. Regarding the other metrics, collisions were found to be considerably more impactful than other disruption causes in terms of performance loss, also known as the loss of resilience (Bruneau et al., 2003), deviation area (Nicholson et al., 2015) or service loss (Chan & Schofer, 2016). Additionally, it was found that performance loss depends more on the maximum impact than on the total duration of a disruption.

Research gap 2: Realization data have not been used to assess the resilience of a railway network for a large and heterogeneous set of disruptions.

To the best of the author's knowledge, this thesis marks the first time that a large and heterogeneous set of disruptions in railways was evaluated by drawing and comparing their resilience curves. Despite the unique characteristics of each disruption, it was shown that the comparison of observed and reported timepoints and statistical analysis of the resilience metrics can provide clues about where to focus effort and resources to improve resilience. Information such as target performance, the start and end time and the steady state could be derived from the realization data itself, provided that details such as the location and the disruption cause are known. However, caution is required to ensure as much as possible that the resilience curve is not contaminated by other disruptions or disturbances, which is necessary to consider the studied disruptions as independent observations. Because of this, one should foresee that the initial sample size may be reduced significantly in data preparation.

Research gap 3: The spatial attributes of a railway network have not been addressed explicitly when studying resilience as a function of time.

Another novelty of this thesis is the study of the disrupted area based on the disruption location and impact type, instead of studying the entire network or a fixed part of the network. The disrupted area for each disruption was determined according to concepts which are specific to disruption management in the Netherlands. This required taking the network structure and realized train paths into account. Although the size of the impact area and the timetable points in the impact area vary with each disruption, the underlying methodology is always the same, which makes it possible to compare resilience curves for disruptions occurring in different parts of the network. Also, by drawing the resilience curve for the entire impact area instead of only the disrupted line, the propagation of disruption effects up to a certain distance away from the disruption location could be represented.

Broader application

Other countries and transport modes

The results from this thesis apply specifically to disruptions in passenger railway traffic in the Netherlands, raising the question that the results might be less relevant for railway networks in other countries or for other transport modes. It is acknowledged that each country is unique in terms of network density, coordination structure, internal business processes, and so on. Although there will likely be similarities in the disruption dynamics among countries, there will also be differences. In countries like Belgium, which rely more heavily on the reactive capacity of traffic controllers and dispatchers, there might be even more types of resilience curves or less obvious differences in terms of the resilience metrics as a result of less standardization. In countries like Germany, which have a more decentralized coordination structure, regional differences in the disruption dynamics might be larger, although it should be noted here that differences between traffic control areas in the Netherlands were not evaluated.

The findings may not be transferable to other transport modes because of the different type of infrastructure, vehicles, business processes and coordination structure. This does not mean however that the resilience evaluation framework cannot be applied to other modes, perhaps after rethinking some of the components in the framework. Non-rail modes such as bus or taxi do not experience the same infrastructure related disruptions that occur in railway networks, and furthermore, they are more flexible in terms of routing because they are not limited to dedicated infrastructure. Similar disruptions as in railways could occur in other rail modes such as subway and light rail, but those networks are typically less interconnected. Hence, a disruption in one place might not necessarily affect the rest of the network. Also, the proportion in which different disruption causes occur will likely be different.

Requirements for recovery functions

A next step in quantifying the resilience of railway networks could be the formulation of analytical recovery functions such as in Cassottana et al. (2021), who developed a recovery function "capable of modeling various performance loss and recovery behaviors, including slow/fast losses of performance followed by faster/slower recoveries". Their function was specified to represent a water distribution system, but recovery functions could be developed to represent other systems as well. A major limitation of the recovery function in Cassottana et al. (2021) with respect to railway resilience is that it does not appear to be capable of handling fluctuating behavior, which was frequently observed in the experiments. Only rarely does the resilience curve go down and up again in a fluent motion, in which case it would classify as the hammock shaped curve. When a recovery function, or a family of recovery functions, is developed to quantify the resilience of a railway network, the function(s) should meet at least the following six requirements:

1. The function(s) shall start and end at the same level of performance.
2. The function(s) shall not take on infeasible levels of performance at any time.
3. The function(s) must be able to handle fluctuations in performance.
4. The function(s) must be able to remain steady for a long period of time.
5. The function(s) must be able to handle both short and long disruptions.
6. The function(s) must be able to handle both low-impact and high-impact disruptions.

6.4. Chapter summary

In this chapter, the results from the experiments were further elaborated upon by highlighting the factors that could play a role in the disruption dynamics, and thus, could affect the resilience of the railway network. Areas for improvement were discussed for disruption management in the Netherlands, while also discussing the preferred shape of the resilience curve, and an outlook to the future was given by discussing how the disruption dynamics might change as

a result of current and future developments. Also, the contributions to railway resilience theory were summarized by reflecting on the identified research gaps.

Answer to subquestion 5

With the knowledge obtained in this chapter, the fifth subquestion is answered.

> **Subquestion 5:** Which insights do quantitative differences and relationships between the resilience metrics bring that may help practitioners evaluate and improve the quality of railway disruption management?

The results from the experiments showed that the disruption dynamics are quite heterogeneous, and that the resilience curve does not necessarily resemble the theoretical shape of the bathtub model. Based on the mean resilience curves and the differences in the resilience metrics among groups, collisions were found to be the most impactful type of single disruption overall, and train defects were found to be the least impactful type of single disruption overall. Collisions stand out in terms of their maximum impact and performance loss, which is why it would be worthwhile to focus more effort on improving the handling of collisions, specifically regarding the swiftness of reporting the collision and the number of trains that are canceled.

Furthermore, it was found that not only train defects are reported late on average, but other types of disruptions are reported late as well, even though quick communication throughout the chain of actors could help prevent secondary delays. To reduce the workload in the first phase, it would be favorable to design predefined solutions for the first phase as well, especially with the prospects of the upcoming frequency increase on busy routes. Since no clear relationship was found between the degradation time and any other resilience metric, it should be stressed that getting quickly into the second phase does not necessarily tell much. Instead, it would be more important to come to a well-advised decision about a structurally feasible traffic plan. The same goes for the transition from second to third phase. Furthermore, it was found that the third phase frequently lasts (much) longer than reported, which may be explained by delays that arise during the restart or a restart plan that proves infeasible. Improvements could be gained with regard to the prognosed end time and the preparation of the restart plan.

7. Conclusions and recommendations

Based on the findings from the literature review, interviews and quantitative data analysis, this chapter presents the conclusions and recommendations from the thesis research. Section 7.1 presents the conclusions by recapitulating the answers to the subquestions and answering the main research question. Section 7.2 presents recommendations with regard to data processing and disruption management practices in ProRail. Section 7.3 discusses the limitations of this research and identifies future research directions from a methodological, practical and scientific perspective.

7.1. Recaps

The answers to the five subquestions defined in Chapter 1 were already given in the last section of each chapter from Chapter 2 onwards. This section briefly recapitulates the answers to those questions after which a general answer to the main research question is given.

Subquestion 1: What can be learned from previous quantitative, data-driven approaches for resilience evaluation of railway networks?

Resilience research exhibits domain-independent characteristics, and so, resilience definitions and metrics are defined in broad terms and could apply to general systems. Performance indicators on the other hand are more domain-specific. Previous data-driven resilience evaluation approaches leave research gaps by failing to study common types of disruptions, not realizing the potential of big data to assess network resilience and not explicitly addressing the spatial attributes of a system. When measuring the (loss of) performance, it is customary to compare actual performance with planned performance and also to aggregate the data. The resilience metrics found in railway resilience literature also appear in general systems literature. Additional metrics describing the nonlinearity of the resilience curve are found in supply chain literature. In case resilience metrics are the main output of a study, statistical analyses are commonly performed to assess their significance or dependence.

Subquestion 2: What is the current state of the practice and quantitative knowledge regarding different types of railway disruptions in the Netherlands?

Each country has its own specific coordination structure and approach to railway disruption management. In the Netherlands, there is a heavy reliance on predefined solutions known as contingency plans, which are designed to function as a revisited timetable. These contingency plans apply to the second resilience phase, although traffic needs to be managed in the first and third phase as well in order to make the transition towards and from the contingency plan, respectively. The plans consider the effects on infrastructure capacity, but not the underlying cause of a disruption. Regarding the state of the practice, the quantitative knowledge about disruptions is limited to the average duration per phase and the total duration. Although the bathtub model describes the disruption dynamics conceptually, the actual evolution of system performance during disruptions is not known, even though the necessary data are available.

Subquestion 3: How can the spatiotemporal effects of disruptions and recovery measures on railway system performance be quantified for the different resilience phases?

The effects of disruptions and recovery measures were described in terms of traffic punctuality, defined as the proportion of punctual to realized train activities, and traffic intensity, defined as the proportion of realized to scheduled train activities. The two indicators are complementary, since punctuality does not include cancellations and traffic intensity does not include

delays. The indicators were combined in a composite performance indicator by means of a weighted sum, where most weight was put on the traffic intensity component. The evolution of system performance during a disruption was calculated as a centered moving average for the entire disrupted area, which is referred to as the impact area. This does not only include the affected line or timetable point, but the entire area in which trains may be canceled in order to manage the disruption. Seven resilience metrics were defined to describe the profile of the resilience curve: degradation time, response time, recovery time, maximum impact, performance loss, degradation profile and recovery profile. The metrics represent the different resilience phases as well as the dimensions of time and performance, where the spatial component is included in the performance dimension. Together, the metrics should account for the multi-dimensional nature of resilience.

Subquestion 4: Which approach should be taken to evaluate railway network resilience for a large and heterogeneous set of disruptions based on traffic realization data?

A resilience evaluation framework was developed which proved successful in evaluating a large and heterogeneous set of disruptions. The evaluation requires at the very least knowledge of the plan time, realization time and location of a train activity; knowledge of the start and end time, location and cause of a disruption; and knowledge of the network structure. The evolution of performance in time can then be calculated, and from those calculations, the resilience curve can be drawn and the resilience metrics can be determined. Group comparisons may be performed to identify differences in the resilience metrics among groups, provided there are no dependencies among the studied disruptions. The specific disruption cause may be chosen here as the variable that defines group membership. Subsequently, regression analysis may be performed to identify relationships between the metrics. The results from these analyses would indicate which disruption types and/or which parts in a disruption to focus on in order to improve resilience. In addition, one resilience metric in particular (the maximum impact) could be monitored in real time in an attempt to predict, and if possible, limit performance loss.

Subquestion 5: Which insights do quantitative differences and relationships between the resilience metrics bring that may help practitioners evaluate and improve the quality of railway disruption management?

The experiments showed that the disruption dynamics are quite heterogeneous, and that the resilience curve does not necessarily follow the theoretical shape of the bathtub model. Differences in the resilience metrics were found to be more prominent than relationships between the metrics. Significant differences exist among disruption causes in terms of the degradation time, response time, recovery time, maximum impact and performance loss. Train defects were identified as the least impactful single disruptions on multiple resilience metrics, where collisions were identified as the most impactful single disruptions on multiple resilience metrics. Therefore, it would be worthwhile to put more effort into improving the handling of collisions, specifically regarding the swiftness of reporting and the number of trains that are canceled in the first phase. Also, it was found that disruptions are reported late on average, even though quick communication would benefit the disruption management process. To reduce the workload in the first phase, it would be favorable to design predefined solutions for the first phase as well. It should be stressed here that getting quickly into the next phase does not necessarily tell much, since no clear relationships were identified between the duration of a specific phase and any other resilience metric. Instead, it would be more important to come to a well-advised decision about a structurally feasible traffic plan. In addition, it was found that the third phase frequently lasts (much) longer than reported. Improvements in the third phase could be gained with regard to the prognosed end time and the preparation of the restart plan.

Main research question: How does the system performance of a railway network develop during disruptions?

As an answer to the main research question, the general conclusion from this thesis 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, despite the fact that all resilience curves result from the same, largely standardized disruption management process. 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 can even be quite unpredictable. Still, a logical explanation for the shape of the curve can usually be found. Among the studied disruption causes, differences exist in terms of the resilience metrics that describe the shape of the resilience curve, and also, some types of resilience curves appear more characteristic of one disruption cause than of another. In terms of traffic intensity (or: cancellations), performance is generally affected the least for train defects and the most for collisions, especially in the beginning. In terms of punctuality (or: delays), performance is generally affected the most for the infrastructure related causes, again, especially in the beginning. The degree in which performance is affected explains a significant part of the cumulative loss of performance, and thus, monitoring the maximum impact in real time could help predict and limit the performance loss.

7.2. Recommendations

Building on the findings from this thesis research, recommendations are made with regard to data processing and disruption management practices in ProRail.

Data processing

With regard to the quality and completeness of traffic realization data and disruption log data in ProRail, the following is recommended:

1. **Make sure the reported timepoints are always available** in case a VSM was applied, so that each disruption can be evaluated in the same way. Specifically, this concerns the timepoints "first phase complete" and "service restored" which formally mark the end of the first and third phase, respectively. Both are dependent on whether or not trains were short-turned, while clearly, short-turning is just one of the management actions. Without short-turning, the resilience curve still has a first, second and third phase for which the start and end time could be identified, perhaps by a change in definition or by applying the methods used in this thesis to derive the timepoints from the resilience curve itself.
2. **Critically review the start and end time of disruptions**, as it was found that disruptions tend to start earlier and end later than reported on average. Regarding the start time, it is advised to reconsider the five minutes that are subtracted as standard from the moment when a notification reaches the control room. Regarding the end time, it is advised to look beyond the first trains that have run again (especially when only few train series are involved), since delays may arise during the restart or the restart plan may prove to be infeasible after all.
3. **Reconsider the definition of the end of the first phase**, since already before the start of the experiments, it was recognized that the formal definition "first phase complete" is usually not an accurate representation of the moment when performance stabilizes and the second phase is reached. Instead, the end of the first phase could simply be defined as the moment when the VSM is applied, which was identified in the interviews as the more common definition, and also, proved to be reasonably accurate on average (especially for train defects).

4. **Improve the coupling of customer hindrance in Sherlock** between a Monitoring ID and a Spoorweb ID. Currently, this coupling occasionally goes wrong, for example when a minor disruption (where no VSM was applied) occurred in the same area shortly before the disruption to which the customer hindrance should apply, or when the boundary points in the Monitoring record differ from the ones in the Spoorweb record.
5. **Improve the VSM assessment in Sherlock** by performing a sanity check for the assessment made by the VLC in Spoorweb. Currently, the assessment is not necessarily accurate and the number of assessments does not always match the number of applied VSMs. The assessment would need to be improved if Sherlock data were to be used to make the comparison between disruptions where a suitable VSM was available and disruptions where no suitable VSM was available.
6. **Make sure realization times and plan times are available** for all train activities at all timetable points, as this would increase the reliability of the performance calculation. Without a realization time, it cannot be assessed whether an activity was punctual or not, and without the most recent plan time, it cannot be ensured that all train paths that are used in determining the impact area are entirely correct. The latter is discussed in more detail in Section 7.3.

Disruption management practices

With regard to disruption management practices in ProRail, the following is recommended:

1. **Raise awareness among practitioners** about the fact that getting quickly into the next phase does not necessarily tell much, and more specifically, that rushing through the first phase does not result in better (or worse) performance during the remainder of the disruption. Instead, it would be more favorable to take the time to come to a well-advised decision about a structurally feasible traffic plan which limits the chance of new delays, as this might lead to a more steady second phase and a smoother and shorter third phase.
2. **Stress the importance of quick information sharing** throughout the chain of actors involved in disruption management. Information sharing should take place regardless of how relevant the information is to the actor who holds the information, so that actors to whom the information is more relevant may act on it. This applies particularly to the start of a disruption, where automating the notification to the control room by a simple solution like the press of a button could initiate the various processes more quickly.
3. **Design predefined solutions for the first phase** to reduce the workload for traffic controllers and dispatchers. The standardization of first-phase measures might leave traffic controllers more time to work towards an optimal VSM. A predefined solution for the first phase could specify the first cancellations and short-turns that are required given a hypothesized disruption location. Since timing is crucial here, the actions could be specified per time window of for example five minutes in a basic hour. Because the exact nature of a disruption is still unclear in the first phase, it would be wise to make the first-phase measure at least as restrictive as a VSM for a full line blockage. Also, the actions should not conflict with the actions in the VSM too much, as this would make the transition from first to second phase more difficult.
4. **Make more accurate prognoses** of the moment when the infrastructure is reclaimed and the train service can be restarted. An accurate prognosis is already a first step towards a feasible restart plan, as this provides TOCs with the necessary constraints to optimize their rolling stock and crew planning until the moment of restart. Furthermore, if the prognosis is advanced, it should be ensured that the TOC has its planning in order before initiating the restart.
5. **Include a restart framework in the VSM more often** and involve TOCs in the design. When a restart framework is available, traffic controllers could invest more time and

expertise in preparing a structurally feasible restart plan while keeping the lines of communication short and efficient. A draft version could still be designed autonomously based on the guidelines in the assessment framework. However, it would be worthwhile to check this version with the TOCs. To improve the usefulness of a restart framework, it is also advised to specify the timeframe in which the train service can be restarted by defining which series can be restarted together and at what rate.

6. **Be more in the lead during the third phase** to ensure a smooth execution of the restart plan. It is true that the prime objective for ProRail is to make the infrastructure available, after which it is the responsibility of the TOCs to run their trains again. Still, a well-executed restart plan with minimal delays is eventually in everyone's benefit and creates better starting conditions in case a new disruption were to occur nearby.
7. **Put more effort into improving the handling of collisions**, as these are relatively common and showed to be the most impactful single disruptions overall. This concerns not only collisions with a person, but other types of collisions as well, especially since it is not always clear from the start if a person is involved or not. Attention may be paid to improving the swiftness of reporting a collision and critically assessing the number of trains that are canceled as emergency measures. If no further improvements can be gained in reducing the impact of collisions, attention should be paid to ways by which to prevent collisions from occurring in the first place, such as better monitoring of persons on or nearby the tracks and making the tracks less accessible.
8. **Conduct more data-driven research** into the evolution of system performance and the quality of disruption management, since resilience will be increasingly important as the railway network grows even busier. ProRail could also take more initiative in this research as the problem owner. The required data are there, although the disruption log data leave room for improvement as discussed earlier in the recommendations.
9. **Invest in mathematical optimization models** for real-time rescheduling solutions. Similar to the solver which is currently being developed to design VSMs for the second phase, models could be developed to design measures for the first and third phase. Separate models would be preferred to allow the transition towards and from the VSM, since the actions in each phase are different, and also to ensure the practicality of the models in case the running time poses constraints on their application. Eventually, the models could be expanded to include the scheduling of replacement transport services.

It is worth noting that the recommendations presented here are mostly in line with the pursued operationalization of the *Koers van VL*, which was last described in 2018 as a vision for 2020. Although the *Koers van VL* is seen as a positive development in which the findings from this thesis can be recognized, it is currently still a vision for the future. Thus, as a final recommendation it is advised to actually realize this transition towards even more predefined, proactive disruption management, so the *Koers van VL* does not remain what it is now: a vision.

7.3. Limitations and future research

Given the scope of this thesis and the followed methodology, there are some limitations to this research which are addressed here. As a result, there are still plenty of opportunities for future research which are discussed next.

Limitations

Although (and partly because) much effort was put into representing the resilience curve accurately for as many disruptions as possible, this research has its limitations. First of all, it was assumed that all delays and cancellations that were observed in the studied impact area during a disruption could be allocated to the disruption, when in fact, not all delays and cancellations are necessarily caused by the disruption itself. There could be other disturbances, events without a logistical record in Sherlock, that occurred simultaneously and thereby contaminated the

resilience curve. Also, maintenance works were not taken into account while these typically reduce network capacity, and thus, constrain the rescheduling options for traffic controllers. As a result, it could be expected that the impact of a disruption is worse in case the disruption occurs near a maintenance site.

Furthermore, not all train paths used in the breadth first search to check for feasible paths may be entirely accurate. Realization times could not be used because those were missing too frequently in order to reconstruct a complete path, and thus, the plan times were used to reconstruct realized train paths. Only realized train paths were considered because those have more accurate plan times which, for NS trains, are called Planning in Tenths of Minutes (PINT). However, if PINT is not available for each timetable point on a particular path, the order of two activities planned in the same minute at different locations (due to a small distance in between) could accidentally be swapped, creating an infeasible path. Another limitation with regard to the traffic realization data is that, as already mentioned in Chapter 4, it is unknown to what extent train paths are erroneously canceled in VOS when a train has in fact completed part of its route. This means the number of cancellations may have been overestimated in some cases.

Another limitation of this research is that of the 1,541 disruptions matching the top five disruption causes, less than one third were preserved to be evaluated as single disruptions with a start and end time that could be derived from the resilience curve. The question is then how representative the results are for the entire population of disruptions. Of the disruptions where a VSM was applied, especially the shorter ones were preserved, since those have a lower chance of being connected than longer disruptions. Nonetheless, longer disruptions were included as well in case they were not connected. However, disruptions lasting multiple days and disruptions starting or ending at nighttime were mostly filtered out, even if they were not connected, because those are more likely to have empty time windows which would turn the resilience curve into a discrete function.

With regard to the statistical analyses, it should be noted that the Python module Pingouin, which was used to perform Welch's ANOVA and the Games-Howell post hoc test, is still under heavy development. For example, as recently as February 2021, an update was released which included a bugfix for an error in the calculation of p-values for the Games-Howell test. Hence, the results of the group comparisons and post hoc tests should be interpreted with caution.

Future research

Methodological improvements

With regard to the methodology, other performance indicators and/or resilience metrics could be used to describe the resilience curve and identify more differences between disruptions of different categories. The variable that defines group membership could be changed from the disruption cause to the impact type, for example to investigate differences between partial and full line blockages. Also, the spatial element could be removed from the performance dimension and presented explicitly on a separate, third axis in a three-dimensional representation of the resilience curve. The resilience curve would then be a cross-section of a resilience plane that describes the evolution of system performance at a certain distance from the disruption location. This could make it easier to assess how fast or how strongly disruption effects propagate through the network. Additionally, it is worth considering other detrending methods than the centered moving average for smoothening the resilience curve, specifically methods that are less sensitive to rapid changes in performance. In the regression analysis, a different kind of regression model such as exponential regression could be applied which might fit the data better, particularly regarding the relationship between performance loss and total duration. The model itself could also be expanded with more dependent variables to reduce the unexplained part of the variation in performance loss.

Practical research directions
With regard to practical research directions in ProRail, it is worth most collecting newer data and repeating the experiments with the new data, considering the current organization of the VGB team and traffic control in ProRail is still relatively new and improvements to disruption management practices have been made continuously in the past years. As a result, the observed differences and relationships could have shifted slightly in the meantime.

In second place, it is worth investigating regional differences in the disruption dynamics, for example between traffic control areas. This might help identify in which areas the network is more resilient. Follow-up research could then try to determine which factors make one part of the network more resilient than another, perhaps building on the factors discussed in Chapter 6, and determine how these factors could be improved in the less resilient parts. Investigating regional differences could also help identify bottlenecks in the network. One potential bottleneck that occurred in the analysis of extreme days is the station in Woerden, which facilitates traffic between major cities in the Randstad. Such bottlenecks might deserve special attention in the design of VSMs, but also in infrastructure development. Regarding regional differences, it is also worth evaluating how different the third phase develops in a regional setting compared to a more nationally oriented restart with a higher share of NS trains. Large differences could require a critical review of the assessment framework, and potentially, targeted improvements in certain TOCs.

In third place, it is worth investigating the effects of (large) maintenance works and freight traffic on the resilience curve, as these were disregarded in the analysis. Perhaps these effects can explain some of the peculiarities in the observed resilience curves. The evolution of performance during disruptions could also be evaluated exclusively for freight traffic. This would ask for a different research design, with different performance indicators (e.g. representing fuel cost and on-time delivery) and a focus on train paths instead of train activities. Since freight trains often do not run according to the original plan, this evaluation could show if (and when) the reactive capacity of traffic controllers is sufficient to manage freight traffic next to the remaining passenger traffic.

In fourth place, it is worth addressing the tradeoff in the number of railway switches, which is a topic that deserves better coordination between asset management and traffic control. On the one hand, reducing the number of switches also reduces the probability of a switch failure occurring anywhere in the network. On the other hand, the number of switch failures that require a VSM is already relatively low compared to the number of train defects, section/signal failures and collisions. Also, switch failures were not found to be that impactful on average, and reducing the number of switches limits the rescheduling options in all other disruptions. Perhaps increased maintenance of switches and other infrastructure elements would be an effective alternative to reduce the number of failures without limiting rescheduling options.

In fifth place, it is worth investigating which parts of the second resilience phase are steady and which parts are not, and whether or not an unsteady second phase is purely the result of human action such as waiting for a mechanic or making inaccurate prognoses. In case of an unsteady second phase, it might need to be reviewed if the applied VSM was the correct one, and if not, which modifications could have helped in reaching a steady second phase.

Scientific research directions
With regard to scientific research into railway resilience, it is worth performing similar data-driven analyses for other countries, while taking the uniqueness of each country into account in terms of network density, coordination structure, business processes etc. For this purpose, the resilience evaluation framework from this thesis could be applied. Although the results might not be directly comparable, they could give an indication of whether or not reactive rescheduling (or “resilient” rescheduling as in Schipper and Gerrits (2018)) improves network

resilience, or if the anticipatory approach followed in the Netherlands should be preferred to achieve resilience. Similar analyses could also be performed for the resilience evaluation of other (public) transport networks after incorporating the mode-specific characteristics.

Another potential research direction is the interaction between punctuality and traffic intensity. This interaction appears to be rather complex, and although it was described in this thesis in terms of the composite performance indicator, it is still not properly understood. Based on the disruptions that were studied in greater detail, such as the example disruption discussed in Chapter 5, no consistent pattern could be identified in this interaction. Particularly, it would be interesting to evaluate what makes punctuality vary so wildly over the course of a disruption on some occasions.

Since connected disruptions were largely excluded from the experiments, it is also worth studying connected disruptions in more detail based on the directions given in Chapter 5, and to see if different types can be identified here as well. A better understanding of connected disruptions could help in finding ways to manage these potential out-of-control situations. Relatively long disruptions, especially those stretching over multiple days, deserve attention in future research as well. For these disruptions, it is worth investigating to what extent performance is affected the next day due to spillover effects and which consequences this has for new disruptions occurring later that day.

As a last scientific research direction, it is worth investigating the expected effects of automatic train operation (ATO) on the disruption dynamics, since ATO removes much of the human element that is so clearly observable in the entire disruption management process. For one, it could be investigated how ATO will change the nature of train defects, as it could help identify and resolve certain failures more easily, but also, introduce new kinds of failures. In addition, the absence of a human driver might affect the handling of other types of disruptions, which is worth studying as well.

Bibliography

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. *Reliability Engineering & System Safety, 153*, 1-14. https://doi.org/10.1016/j.ress.2016.03.015

Altay, N., & Green, W. G. (2006). OR/MS research in disaster operations management. *European Journal of Operational Research, 175*(1), 475-493. https://doi.org/10.1016/j.ejor.2005.05.016

Azad, N., Hassini, E., & Verma, M. (2016). Disruption risk management in railroad networks: An optimization-based methodology and a case study. *Transportation Research Part B: Methodological, 85*, 70-88. https://doi.org/10.1016/j.trb.2016.01.001

Baroud, H., Barker, K., Ramirez-Marquez, J. E., & Rocco, C. M. (2014). Importance measures for inland waterway network resilience. *Transportation Research Part E: Logistics and Transportation Review, 62*, 55-67. https://doi.org/10.1016/j.tre.2013.11.010

Bashan, A., Bartsch, R., Kantelhardt, J. W., & Havlin, S. (2008). Comparison of detrending methods for fluctuation analysis. *Physica A: Statistical Mechanics and its Applications, 387*(21), 5080-5090. https://doi.org/10.1016/j.physa.2008.04.023

Bešinović, N. (2020). Resilience in railway transport systems: a literature review and research agenda. *Transport Reviews, 40*(4), 457-478. https://doi.org/10.1080/01441647.2020.1728419

Bevilacqua, M., Ciarapica, F. E., & Marcucci, G. (2017). Supply Chain Resilience Triangle: The Study and Development of a Framework. *International Journal of Economics and Management Engineering, 11*(8), 2046-2053. https://doi.org/10.5281/zenodo.1131597

Brandon-Jones, E., Squire, B., Autry, C. W., & Petersen, K. J. (2014). A Contingent Resource-Based Perspective of Supply Chain Resilience and Robustness. *Journal of Supply Chain Management, 50*(3), 55-73. https://doi.org/10.1111/jscm.12050

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. *Earthquake Spectra, 19*(4), 733-752. https://doi.org/10.1193%2F1.1623497

Büchel, B., Spanninger, T., & Corman, F. (2020). Empirical dynamics of railway delay propagation identified during the large-scale Rastatt disruption. *Scientific Reports, 10*, 18584. https://doi.org/10.1038/s41598-020-75538-z

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. *Transportation Research Part B: Methodological, 63*, 15-37. https://doi.org/10.1016/j.trb.2014.01.009

Cao, S., & Rhinehart, R. R. (1995). An efficient method for on-line identification of steady state. *Journal of Process Control, 5*(6), 363-374. https://doi.org/10.1016/0959-1524(95)00009-F

Carvalho, H., & Cruz-Machado, V. (2011). Integrating Lean, Agile, Resilience and Green Paradigms in Supply Chain Management (LARG_SCM). In P. Li (Ed.), *Supply Chain Management*. InTech.

Cassottana, B., Aydin, N. Y., & Tang, L. C. (2021). Quantitative Assessment of System Response during Disruptions: An Application to Water Distribution Systems. *Journal of Water Resources Planning and Management, 147*(3), 04021002. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001334

Castiglioni, P., & Di Rienzo, M. (2004). How to check steady-state condition from cardiovascular time series. *Physiological Measurement. 25*(4), 985-996. https://doi.org/10.1088/0967-3334/25/4/016

Cats, O., & Jenelius, E. (2014). Dynamic Vulnerability Analysis of Public Transport Networks: Mitigation Effects of Real-Time Information. *Networks and Spatial Economics, 14*, 435-463. https://doi.org/10.1007/s11067-014-9237-7

Chan, R., & Schofer, J. L. (2016). Measuring Transportation System Resilience: Response of Rail Transit to Weather Disruptions. *Natural Hazards Review, 17*(1), 05015004. https://doi.org/doi:10.1061/(ASCE)NH.1527-6996.0000200

Chen, L., & Miller-Hooks, E. (2012). Resilience: An Indicator of Recovery Capability in Intermodal Freight Transport. *Transportation Science, 46*(1), 109-123. https://doi.org/10.1287/trsc.1110.0376

Cimellaro, G. P., Reinhorn, A. M., & Bruneau, M. (2010). Seismic resilience of a hospital system. *Structure and Infrastructure Engineering, 6*(1-2), 127-144. https://doi.org/10.1080/15732470802663847

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Dalheim, Ø. Ø., & Steen, S. (2020). A computationally efficient method for identification of steady state in time series data from ship monitoring. *Journal of Ocean Engineering and Science, 5*(4), 333-345. https://doi.org/10.1016/j.joes.2020.01.003

Dekker, M. M., & Panja, D. (2021). Cascading dominates large-scale disruptions in transport over complex networks. *PLOS ONE, 16*(1), e0246077. https://doi.org/10.1371/journal.pone.0246077

Dekker, M. M., Van Lieshout, R. N., Ball, R. C., Bouman, P. C., Dekker, S. C., Dijkstra, H. A., Goverde, R. M. P., Huisman, D., Panja, D., Schaafsma, A. A. M., & Van den Akker, M. (2021). A next step in disruption management: combining operations research and complexity science. *Public Transport*. https://doi.org/10.1007/s12469-021-00261-5

De-Los-Santos, A., Laporte, G., Mesa, J. A., & Perea, F. (2012). Evaluating passenger robustness in a rail transit network. *Transportation Research Part C: Emerging Technologies, 20*(1), 34-46. https://doi.org/10.1016/j.trc.2010.09.002

D'Lima, M., & Medda, F. (2015). A new measure of resilience: An application to the London Underground. *Transportation Research Part A: Policy and Practice, 81*, 35-46. https://doi.org/10.1016/j.tra.2015.05.017

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.

Draper, D. (1988). Rank Based Robust Analysis of Linear Models. I. Exposition and Review. *Statistical Science, 3*(2), 239-257. https://doi.org/10.1214/ss/1177012915

European Rail Research Advisory Council. (2020). *Rail Strategic Research and Innovation Agenda*. Click Click Graphics, https://shift2rail.org/wp-content/uploads/2020/12/RAIL-Strategic-Research-and-Innovation-Agenda-2020-_FINAL_dec2020.pdf

Evans, I. (2011). Railway disruption recovery: lessons from airlines. *WIT Transactions on Modelling and Simulation, 51*, 681-692. https://doi.org/10.2495/CMEM110601

Ghaemi, N., Cats, O., & Goverde, R. M. P. (2017). Railway disruption management challenges and possible solution directions. *Public Transport, 9*, 343-364. https://doi.org/10.1007/s12469-017-0157-z

Ghaemi, N., Cats, O., & Goverde, R. M. P. (2018). Macroscopic multi-station short-turning model in case of complete blockages. *Transportation Research Part C: Emerging Technologies, 89*, 113-132. https://doi.org/10.1016/j.trc.2018.02.006

Gonçalves, L. A. P. J., & Ribeiro, P. J. G. (2020). Resilience of urban transportation systems. Concept, characteristics, and methods. *Journal of Transport Geography, 85*, 102727. https://doi.org/10.1016/j.jtrangeo.2020.102727

Goverde, R. M. P., Corman, F., & D'Ariano, A. (2013). Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions. *Journal of Rail Transport Planning & Management, 3*(3), 78-94. https://doi.org/10.1016/j.jrtpm.2013.12.001

Goverde, R. M. P., & Hansen, I. A. (2013). Performance Indicators for Railway Timetables. *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings*, 301-306. https://doi.org/10.1109/ICIRT.2013.6696312

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. *European Journal of Epidemiology, 31*, 337-350. https://dx.doi.org/10.1007%2Fs10654-016-0149-3

Henry, D., & Ramirez-Marquez, J. E. (2012). Generic metrics and quantitative approaches for system resilience as a function of time. *Reliability Engineering & System Safety, 99*, 114-122. https://doi.org/10.1016/j.ress.2011.09.002

Hosseini, S., Barker, K., & Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. *Reliability Engineering & System Safety, 145*, 47-61. https://doi.org/10.1016/j.ress.2015.08.006

Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons, https://doi.org/10.1002/0471725250

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering, 9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Ivanov, D., Sokolov, B., & Dolgui, A. (2014). The Ripple effect in supply chains: trade-off 'efficiency-flexibility-resilience' in disruption management. *International Journal of Production Research, 52*(7), 2154-2172. https://doi.org/10.1080/00207543.2013.858836

Jafino, B. A., Kwakkel, J., & Verbraeck, A. (2020). Transport network criticality metrics: a comparative analysis and a guideline for selection. *Transport Reviews, 40*(2), 241-264. https://doi.org/10.1080/01441647.2019.1703843

Jalali, S., & Wohlin, C. (2012). Systematic Literature Studies: Database Searches vs. Backward Snowballing. *Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 29-38. https://doi.org/10.1145/2372251.2372257

Janić, M. (2015). Reprint of "Modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event". *Transportation Research Part A: Policy and Practice, 81*, 77-92. https://doi.org/10.1016/j.tra.2015.07.012

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. https://doi.org/10.1007/s11116-018-9875-6

Jespersen-Groth, J., Potthoff, D., Clausen, J., Huisman, D., Kroon, L., Maróti, G., & Nielsen, M. N. (2009). Disruption Management in Passenger Railway Transportation. In R. K. Ahuja, R. H. Möhring, & C. D. Zaroliagis (Eds.), *Robust and Online Large-Scale Optimization*. (pp. 399-421). Springer.

Jin, J. G., Tang, L. C., Sun, L., & Lee, D.-H. (2014). Enhancing metro network resilience via localized integration with bus services. *Transportation Research Part E: Logistics and Transportation Review, 63*, 17-30. https://doi.org/10.1016/j.tre.2014.01.002

Kelly, J. D., & Hedengren, J. D. (2013). A steady-state detection (SSD) algorithm to detect non-stationary drifts in processes. *Journal of Process Control, 23*(3), 326-331. https://doi.org/10.1016/j.jprocont.2012.12.001

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

Liao, W., Tordeux, A., Seyfried, A., Chraibi, M., Drzycimski, K., Zheng, X., & Zhao, Y. (2016). Measuring the steady state of pedestrian flow in bottleneck experiments. *Physica A: Statistical Mechanics and its Applications, 461*, 248-261. https://doi.org/10.1016/j.physa.2016.05.051

Liu, H. (2015). *Comparing Welch's ANOVA, a Kruskal-Wallis test and traditional ANOVA in case of Heterogeneity of Variance* [Master's thesis, Virginia Commonwealth University]. https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=5026&context=etd

Luo, Y., Li, Z., & Wang, Z. (2009). Adaptive CUSUM control chart with variable sampling intervals. *Computational Statistics and Data Analysis, 53*(7), 2693-2701. https://doi.org/10.1016/j.csda.2009.01.006

Macdonald, J. R., Zobel, C. W., Melnyk, S. A., & Griffis, S. E. (2018). Supply chain risk and resilience: theory building through structured experiments and simulation. *International Journal of Production Research, 56*(12), 4337-4355. https://doi.org/10.1080/00207543.2017.1421787

Madni, A. M., Erwin, D., & Sievers, M. (2020). Constructing Models for System Resilience: Challenges, Concepts, and Formal Methods. *Systems, 8*(3). https://doi.org/10.3390/systems8010003

Malandri, C., Fonzone, A., & Cats, O. (2018). Recovery time and propagation effects of passenger transport disruptions. *Physica A: Statistical Mechanics and its Applications, 505*, 7-17. https://doi.org/10.1016/j.physa.2018.03.028

Mattsson, L.-S., & Jenelius, E. (2015). Vulnerability and resilience of transport systems – A discussion of recent research. *Transportation Research Part A: Policy and Practice, 81*, 16-34. https://doi.org/10.1016/j.tra.2015.06.002

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56-61. https://doi.org/10.25080/Majora-92bf1922-00a

Melnyk, S. A., Closs, D. J., Griffis, S. E., Zobel, C. W., & Macdonald, J. R. (2014). Understanding Supply Chain Resilience. *Supply Chain Management Review, 18*, 34-41. https://www.researchgate.net/publication/285800059_Understanding_supply_chain_resilience

Mertens, W., Pugliese, A., & Recker, J. (2017). *Quantitative Data Analysis: A Companion for Accounting and Information Systems Research*. Springer. https://doi.org/10.1007/978-3-319-42700-3

Metselaar, D. (2021, January 5). *Coronacrisis leidt tot halvering check-ins OV in 2020*. OVPro. https://www.ovpro.nl/special/2021/01/05/coronacrisis-leidt-tot-halvering-check-ins-ov-in-2020/?gdpr=accept

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. *International Journal of Production Research, 53*(22), 6736-6751. https://doi.org/10.1080/00207543.2015.1057296

Nicholson, G. L., Kirkwood, D., Roberts, C., & Schmid, F. (2015). Benchmarking and evaluation of railway operations performance. *Journal of Rail Transport Planning & Management, 5*(4), 274-293. https://doi.org/10.1016/j.jrtpm.2015.11.004

Ouyang, M., Dueñas-Osorio, L., & Min, X. (2012). A three-stage resilience analysis framework for urban infrastructure systems. *Structural Safety, 36-37*, 23-31. https://doi.org/10.1016/j.strusafe.2011.12.004

Parkinson, H. J., & Bamford, G. (2017). A journey into railway digitisation. *Stephenson Conference: Research for Railways 2017*, 333-340. https://www.researchgate.net/publication/320352839_A_journey_into_railway_digitisation

Pettit, T. J., Croxton, K. L., & Fiksel, J. (2019). The Evolution of Resilience in Supply Chain Management: A Retrospective on Ensuring Supply Chain Resilience. *Journal of Business Logistics, 40*(1), 56-65. https://doi.org/10.1111/jbl.12202

Ponomarov, S. Y., & Holcomb, M. C. (2009). Understanding the concept of supply chain resilience. *The International Journal of Logistics Management, 20*(1), 124-143. https://doi.org/10.1108/09574090910954873

ProRail. (2018). *Programma Hoogfrequent Spoorvervoer (PHS)* [Map]. https://www.prorail.nl/siteassets/homepage/programmas/documenten/phs-algemene-kaart-1.pdf

ProRail. (2020). *Jaarverslag 2019*. https://www.prorail.nl/siteassets/homepage/over-ons/documenten/jaarverslag-2019-prorail.pdf

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. https://doi.org/10.3850/978-981-11-2724-3_0235-cd

Rijden de Treinen. (n.d.). *Statistieken*. Retrieved March 16, 2021, from https://www.rijdendetreinen.nl/statistieken

Rodrigue, J.-P. (2020). *The Geography of Transport Systems* (5th Edition). Routledge. https://transportgeography.org/

Schipper, D., & Gerrits, L. (2018). Differences and similarities in European railway disruption management practices. *Journal of Rail Transport Planning & Management, 8*(1), 42-55. https://doi.org/10.1016/j.jrtpm.2017.12.003

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference,* 92-98. http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf

Simons, R. (2019). *The influence of railway signalling characteristics on resilience* [Master's thesis, Delft University of Technology]. https://repository.tudelft.nl/islandora/object/uuid:0e76a919-3d02-4282-a56b-dd20a94b61fa

Spiegler, V. L. M., Naim, M. M., & Wikner, J. (2012). A control engineering approach to the assessment of supply chain resilience. *International Journal of Production Research, 50*(21), 6162-6187. https://doi.org/10.1080/00207543.2012.710764

Theil, H. (1971). *Principles of Econometrics*. Wiley. https://archive.org/details/principlesofecon0000thei/page/n5/mode/2up

Tsuchiya, S., Tatano, H., & Okada, N. (2007). Economic Loss Assessment due to Railroad and Highway Disruptions. *Economic Systems Research, 19*(2), 147-162. https://doi.org/10.1080/09535310701328567

Tukamuhabwa, B. R., Stevenson, M., Busby, J., & Zorzini, M. (2015). Supply chain resilience: definition, review and theoretical foundations for further study. *International Journal of Production Research, 53*(18), 5592-5623. http://dx.doi.org/10.1080/00207543.2015.1037934

Uday, P., & Marais, K. (2015). Designing Resilient Systems-of-Systems: A Survey of Metrics, Methods, and Challenges. *Systems Engineering, 18*(5), 491-510. https://doi.org/10.1002/sys.21325

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software, 3*(31), 1026. https://doi.org/10.21105/joss.01026

Van Aken, S., Bešinović, N., & Goverde, R. M. P. (2017). Designing alternative railway timetables under infrastructure maintenance possessions. *Transportation Research Part B: Methodological, 98*, 224-238. https://doi.org/10.1016/j.trb.2016.12.019

Van Hoek, R. (2020). Research opportunities for a more resilient post-COVID-19 supply chain – closing the gap between research findings and industry practice. *International Journal of Operations & Production Management, 40*(4), 341-355. https://doi.org/10.1108/IJOPM-03-2020-0165

Van Wee, B., & Banister, D. (2016). How to Write a Literature Review Paper? *Transport Reviews, 36*(2), 278-288. https://doi.org/10.1080/01441647.2015.1065456

Veelenturf, L. P. (2014). *Disruption Management in Passenger Railways* [Doctoral thesis, Erasmus University]. https://repub.eur.nl/pub/77155

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

Wong, A., Tan, S., Chandramouleeswaran, K. R., & Tran, H. T. (2020). Data-driven analysis of resilience in airline networks. *Transportation Research Part E: Logistics and Transportation Review, 143*, 102068. https://doi.org/10.1016/j.tre.2020.102068

Woodburn, A. (2019). Rail network resilience and operational responsiveness during unplanned disruption: A rail freight case study. *Journal of Transport Geography, 77*, 59-69. https://doi.org/10.1016/j.jtrangeo.2019.04.006

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*(12), 2141-2155. https://doi.org/10.1080/00949655.2010.520163

Zhou, Y., Wang, J., & Yang, H. (2019). Resilience of Transportation Systems: Concepts and Comprehensive Review. *IEEE Transactions on Intelligent Transport Systems, 20*(12), 4262-4276. https://doi.org/10.1109/TITS.2018.2883766

Zhu, Y., & Goverde, R. M. P. (2021). Dynamic railway timetable rescheduling for multiple connected disruptions. *Transportation Research Part C: Emerging Technologies, 125*, 103080. https://doi.org/10.1016/j.trc.2021.103080

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. *Transportation Research Record: Journal of the Transportation Research Board, 2599*(1), 70-80. https://doi.org/10.3141%2F2599-09

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. *Transportation Research Record: Journal of the Transportation Research Board, 2604*(1), 9-18. https://doi.org/10.3141%2F2604-02

Zilko, A. A., Kurowicka, D., & Goverde, R. M. P. (2016). Modeling railway disruption lengths with Copula Bayesian Networks. *Transportation Research Part C: Emerging Technologies, 68*, 350-368. https://doi.org/10.1016/j.trc.2016.04.018

Zobel, C. W. (2011). Representing perceived tradeoffs in defining disaster resilience. *Decision Support Systems, 50*(2), 394-403. https://doi.org/10.1016/j.dss.2010.10.001

Appendices

A. Research paper

This appendix contains the research paper which summarizes the problem, approach, methods and most important scientific results from this thesis report.

A Data-Driven Approach for Evaluating the Resilience of Railway Networks

Max J. Knoester, Nikola Bešinović[a], Amir P. 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 Staff, ProRail, Utrecht, The Netherlands*

ARTICLE INFO

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

ABSTRACT

Disruptions occur frequently in railway networks, requiring adjustments to the timetable, rolling stock planning and crew planning while causing delays and cancellations. Although the evolution of system performance during a disruption can be visualized in the resilience curve, not much is known about performance during disruptions or the extent to which the curve applies in practice. The limited quantitative knowledge about the resilience of railway networks makes it hard to design appropriate recovery measures. In this paper, a data-driven evaluation approach is presented to make an ex post assessment of the resilience of railway networks. Several resilience metrics are extracted from literature and two new resilience metrics are introduced. Using historical traffic realization data, resilience curves are reconstructed for a large and heterogeneous set of single disruptions and are quantified in terms of the resilience metrics. Among others, the values of the resilience metrics are compared across disruptions of different causes using Welch's ANOVA and the Games-Howell test. The approach is applied to a case study of the Dutch railway network, with a focus on the five most common disruption causes. The results of the case study show that there is significant heterogeneity in the shape of the resilience curve, even within disruptions of the same cause. Train defects are found to be the least impactful disruptions on multiple resilience metrics, while collisions are found to be the most impactful disruptions on multiple resilience metrics. The successful application of the approach shows that it can be used by practitioners to assess which types and which parts of disruptions deserve attention to improve disruption management practices, and thus, improve resilience.

1 Introduction

The Dutch railway network is known as the busiest in Europe (ACM, 2019), with approximately 1.3 million passenger trips and 148 million ton kilometers of freight transport every day. Under normal conditions, trains arrive and depart according to the timetable and only minor variations in the train service are observed, which 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). Where 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 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). When this is the case, primary delays will have caused extensive secondary delays, and imbalances in the available rolling stock and crew 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 Figure 1. The resilience curve shows how performance first degrades and eventually recovers. Three phases are distinguished in the curve: the first and the third phase are transition phases, whereas the second phase represents disrupted but stable system behavior. The resilience curve is sometimes referred to as the bathtub model (Ghaemi et al., 2017), since it is presumed to resemble a bathtub.

Traffic control during disruptions is also referred to as rescheduling and is commonly performed by the infrastructure manager. Rescheduling is anticipation-based in case recovery measures are predefined. However, if rescheduling happens mostly in real time, tailor-made solutions have to be made for each disruption, which places 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, traffic control is the task of 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 (*versperringsmaatregel, VSM*) is applied which provides a revisited timetable for the second phase. In the second phase, the execution of the VSM is monitored and the cause of the disruption is resolved. Meanwhile, a restart plan is

prepared to resume the train service according to the original timetable. When the plan is approved and the infrastructure is reclaimed, the restart can be initiated. In the third phase, the execution of the restart 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).

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, which is why this 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?

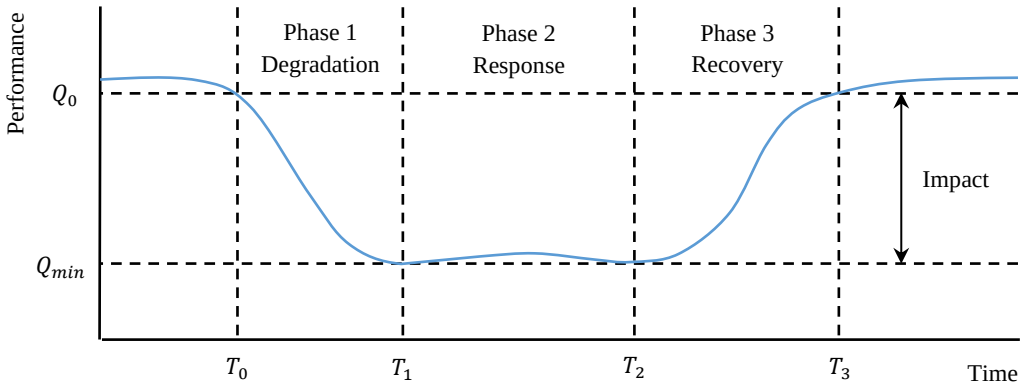


Figure 1. Schematic illustration of the theoretical resilience curve.

In this paper, a data-driven evaluation approach is presented to make an ex post assessment of the resilience of railway networks using historical traffic realization data and disruption data. Several resilience metrics that quantify the shape of the resilience curve are extracted from literature and two new resilience metrics are introduced: the degradation profile and the recovery profile. For a large and heterogeneous set of common disruptions, the resilience curve is reconstructed and the phases in the resilience curve are identified. Each resilience curve is described in terms of the resilience metrics. The values of the metrics are evaluated in statistical analyses to identify differences and similarities among disruptions of different causes. The proposed approach is data-driven, as it uses historical traffic realization data which most importantly specify the plan time and realization time of a train activity. 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 evaluation approaches (e.g. topological, optimization-based, simulation-based), the benefits of this data-driven approach are that it removes the need to model the traffic conditions in the network explicitly, and that it allows a direct comparison between what practitioners believe to be true and what happens in reality.

This study contributes to the field of transport and railway resilience research from the following perspectives. First of all, a new data-driven approach is presented, which is summarized in a resilience evaluation framework, to ex post assess the resilience of railway networks. A composite performance indicator is defined which is capable of representing delays and cancellations simultaneously. Two new resilience metrics are introduced and a method is proposed for identifying the resilience phases from the resilience curve. The final contribution consists of empirical testing of the proposed evaluation approach on a 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 evaluation framework and discusses the methodology for the resilience evaluation. Section 4 provides the case study description and presents the results of empirical testing of the proposed evaluation approach on the case study. Section 5 provides a more thorough discussion on the results. Section 6 concludes the paper and presents future research directions.

2 Literature review

Numerous studies have investigated the resilience of railway networks. A review 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 may be formulated as follows. Resilience is the ability of a 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. To provide a thorough literature review on the quantification of resilience in previous studies, we first review the existing resilience evaluation

approaches in Section 2.1. Different types of approaches include topological, optimization-based, simulation-based and data-driven. Here, we focus on data-driven approaches in railway transport and other transport modes including subway, taxi and air transport. Next, in Section 2.2 we review existing resilience metrics, which are used later for selecting appropriate metrics that describe the shape of the resilience curve. Section 2.3 presents the existing research gaps in railway resilience research.

2.1 Evaluation 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.

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. 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 in 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 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.

To summarize, previous data-driven studies on the resilience of transport systems mainly examined single, large-scale disruptive events which are relatively uncommon. Only Ren et al. (2020) and Wong et al. (2020) studied a large number of disruptions, but they did not do so 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 shape of the curve quantitatively. Assuming that the exact mathematical function of the curve is not known, one way 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), which is why 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, recovery rate, deterioration rate, initial impact, maximum impact and minimum performance (e.g. Nicholson et al., 2015; Janić, 2018). Minimum performance may also be referred to as residual functionality (Cimellaro et al., 2010). Recovery time is in fact 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.

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)

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 the resilience curve consists 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 from supply chain literature, 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 large-scale disruptions, while in reality, disruptions of a smaller scale such as switch failures and train defects occur frequently. Yet, these disruptions have not been subject to resilience research.

Second, 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.

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

3 Methodology

Based on theoretical and practical insights, a data-driven approach is developed to ex post evaluate the resilience of railway networks given a large number of disruptions of different types. The approach is generic and can work for different performance indicators, resilience metrics and categorizations of disruption types. The approach is structured in the resilience evaluation framework presented in Figure 2. The framework has been divided into three parts: input, processing and output. In short, the framework works as follows. The evaluation starts with the collection of traffic realization data, disruption log data and network data from a database. Traffic realization data should include at least the plan time, realization time and location of a train activity. 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 for 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 meant to identify differences and similarities among disruptions of different types. The resilience metrics and the test statistics are the quantitative output of the resilience evaluation. The resilience curves, which can be drawn directly from the performance measurements, are the graphic output of the resilience evaluation. Apart from plotting individual resilience curves, one could also identify different types of resilience curves and the mean and median resilience curve per disruption type.

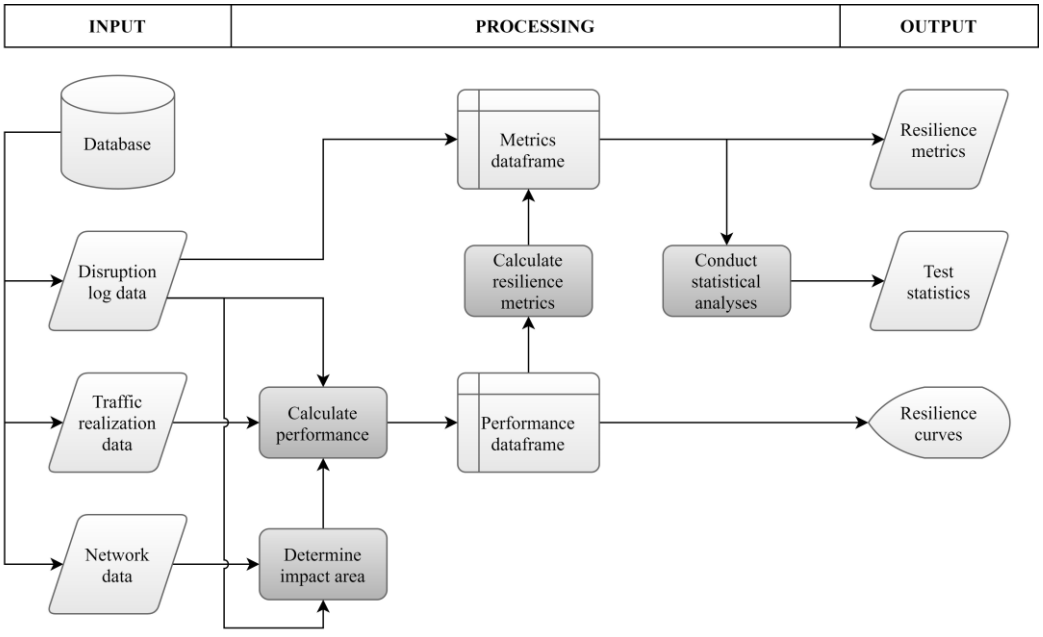


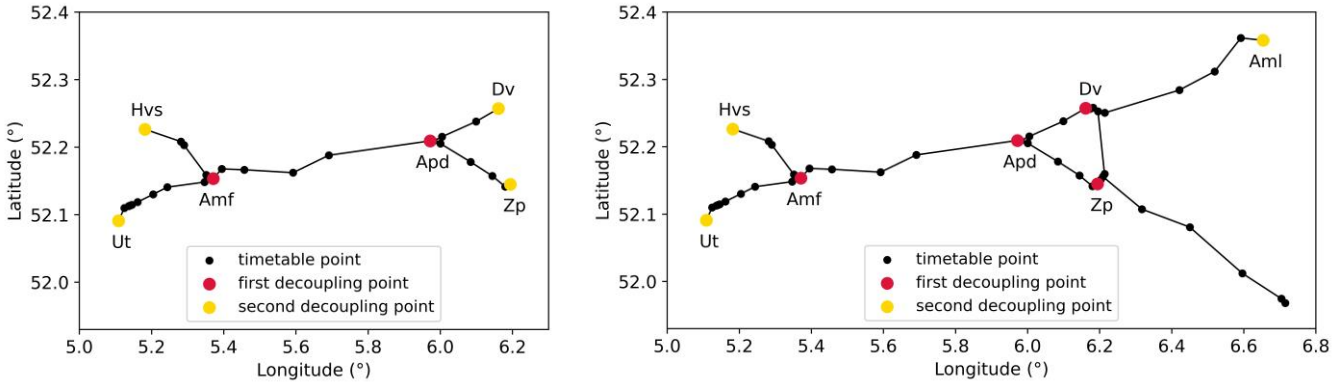
Figure 2. The resilience evaluation framework.

The process blocks in the resilience evaluation 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 the 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. The first impact area is bounded by the first intercity decoupling points from the disruption location; the second impact area is bounded by the 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 in the first and second impact area. Thus, the first and second impact area are identified as the appropriate area to consider. The size of this area depends not only on the location in the network, but also on the type of impact. This is illustrated in Figure 3, which shows the first and second impact area of a hypothetical disruption for a partial or full line blockage in Figure 3 (a) and for a full timetable point outage in Figure 3 (b).

To determine the impact area of each disruption given the disruption location, a modified breadth first search algorithm is developed. Breadth first search, first described 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. Basically, the algorithm starts from a source node, referred to as the start vertex, and visits any adjacent, unvisited vertices until none are left. Several constrains need to be imposed on the basic algorithm in order to use it in a railway context. For details on the algorithm, the interested reader may refer to Knoester (2021).



(a) Partial/full line blockage

(b) Full timetable point outage

Figure 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).

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 scheduled train activities in a given time period. While the traffic intensity indicator accounts for cancellations, it does not account for delays. Therefore, we consider the traffic punctuality as well. Traffic punctuality is defined as the proportion of punctual train activities (with a delay less than three minutes) relative to 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/T):

$$Q = \left((1-\lambda)\frac{P}{R} + \lambda\frac{R}{T} \right) \cdot 100\% \tag{1}$$

where P is the number of punctual train activities, R is the number of realized train activities, T is the number of scheduled train activities and λ is the normalized performance weight. A weight $\lambda < 0.5$ puts more emphasis on punctuality, whereas a weight $\lambda > 0.5$ puts more emphasis on traffic intensity. Although the composite indicator is somewhat abstract and less easy to communicate

than a single indicator, it does have its benefits. For one, it makes it possible to measure delays and cancellations simultaneously. It also smoothens fluctuations in these indicators, which are most prominent in the punctuality component. Furthermore, 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. Lastly, 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 traffic intensity were to be studied, these effects would not be observed.

Performance at each time instant is calculated according to Equation (1) as a centered moving average over a time period of 30 minutes and with a step size of one minute. A smaller time period would be better able to show the dynamics in system performance, but would also make the resilience curve more difficult to analyze. Calculations are 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 metrics are defined as follows:

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

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

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

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

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

$$DP = \sum_{j=1}^{m} \Big(f\big(t_j\big) - Q\big(t_j\big)\Big) \qquad T_0 \le t_j \le T_1 \tag{7}$$

$$RP = \sum_{k=1}^{s} \big(g(t_k) - Q(t_k)\big) \qquad T_2 \le t_k \le T_3 \tag{8}$$

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; $f(t)$ and $g(t)$ are the linear degradation and recovery function, respectively; n is the total number of time intervals; 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 though, a higher value indicates a stronger disruptive effect, and thus, a less resilient network.

In order to calculate the resilience metrics, it is necessary to identify the timepoints $T_0 \dots T_3$ which are shown in Figure 1. All timepoints are derived from the resilience curve itself. T_0 is defined as the last moment before the reported start of the disruption when performance is still above target, and T_3 is defined as the first moment after the restart is initiated when performance is above target again. Target performance is defined here as the average networkwide performance during the day, measured over a number of relatively quiet days throughout the year. Assuming an approximately steady second phase, T_1 and T_2 are determined by applying a steady state detection algorithm. A steady state detection algorithm is generally used to identify the steady parts of time series data. The approach taken in Dalheim and Steen (2020), which involves fitting a regression model to consecutive, overlapping time windows, is taken as a starting point for the algorithm. The algorithm is modified to account for the dynamic nature of railway system performance. For 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 and similarities in the resilience metrics among disruptions of different types. Knowing where and how large these differences are could help improve resilience in specific parts of certain types of disruptions. 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. Because the distributions of the resilience metrics violate the assumptions of one-way ANOVA as well as the Kruskal-Wallis test, we use Welch's ANOVA for the group comparisons. This method is similar to one-way ANOVA, but it 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 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, 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 only reveals whether a difference exists among groups, but it does not tell 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 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 disruption dynamics. First, Section 4.1 defines the scope of the case study. Section 4.2 presents the detailed evaluation of an arbitrary disruption to illustrate the working of the resilience evaluation framework. Section 4.3 presents the different types of resilience curves that are observed. 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 rail, intercity rail, high-speed rail and international rail. The study area includes the entire Dutch railway network with the exception of traffic control area Kijfhoek, since Kijfhoek exclusively handles freight traffic. In total, 2,152 disruptions occurred in this area and time period for which a capacity reallocation and usually also a VSM 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. 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 contamination of the resilience curve, we focus on single disruptions, which are defined as disruptions that are not related to other nearby disruption. If a disruption cannot be treated as a single disruption, then it is connected. Connected disruptions are identified as two or more disruptions that have overlapping time periods and that have at least one timetable point in their impact area in common. Note that this does not necessarily imply a causal relationship between connected disruptions. In addition, we exclude disruptions on black days (i.e. days when traffic punctuality is below 75% and/or traffic intensity is below 90% network-wide) and near-black days; disruptions with an impact area of less than six timetable points; disruptions with a reported duration longer than ten hours; and disruptions with a missing time entry for the moment when the restart was initiated. This leaves 706 disruptions of the initial 1,541 disruptions that match the top five causes, which is primarily due to excluding connected disruptions.

For the remaining disruptions, it must be ensured that the resilience curve can be described properly so that calculating the resilience metrics is justified. Therefore, we also exclude disruptions for which performance remained above target for the entire reported duration; disruptions for which the start time cannot be identified, meaning performance was already below target at least 60 minutes before the reported start of the disruption; disruptions for which the end time cannot be identified, meaning performance was still below target at least 180 minutes after the reported end of the disruption; disruptions for which an empty time window is encountered in the performance calculation; and disruptions for which a steady state cannot be identified.

How many disruptions are excluded based on the former conditions depends on the way performance evolved, and thus, on the value of the performance weight λ. We consider $\lambda = 0.67$ as a starting point, which puts twice the weight on traffic intensity compared to punctuality. At this value, the resilience curve is relatively well behaved while the punctuality component is dominant enough so that changes in punctuality may be observed. The remaining number of disruptions for $\lambda = 0.67$ is 445 out of 706. Other values of λ yield similar results. An overview of the original and the remaining number of disruptions for the studied disruption causes is presented in Table 2.

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

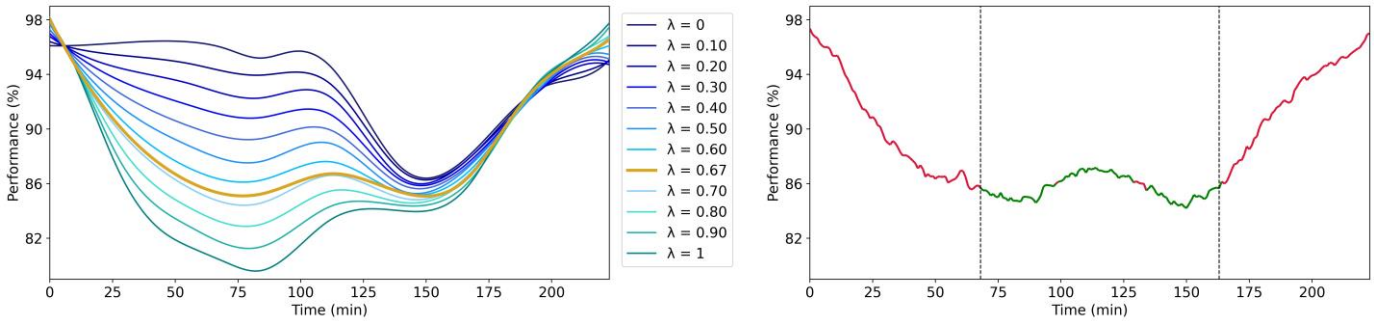
Disruption cause	Number of disruptions in the original dataset	Number of disruptions potentially studied as single disruptions	Number of disruptions studied as single disruptions for the specified performance weight ($\lambda = 0.67$)
Train defect	742	346	202
Section/signal failure	306	141	97
Collision	275	146	96
Switch failure	153	47	34
Overhead line failure	65	26	16
Total	1,541	706	445

4.2 Characteristics of the resilience curve for an example case

An arbitrary disruption is selected to illustrate the working of the resilience evaluation framework up until the statistical analysis. The example case is a collision that occurred between Putten and Nunspeet on April 4, 2019. The collision occurred at 14:43 and resulted in a full line blockage of the double track line. 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. Performance is calculated firstly for different values of the performance weight. The resulting resilience curves are presented in Figure 4 (a) as cubic splines fitted to the actual performance measurements. It is observed that traffic intensity ($\lambda = 1$) dropped quickly at the start of the disruption, while punctuality ($\lambda = 0$) remained relatively stable until approximately 100 minutes into the disruption. The moment when punctuality eventually dropped matches the moment when the evacuation of the stranded passengers and retrieval of the damaged train began. Meanwhile, traffic intensity had already partly recovered. Performance recovered similarly for the different values of the performance weight, which indicates that the restart was executed 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 the initial choice of λ, which is why $\lambda = 0.67$ is maintained throughout the experiments.

The steady state detection algorithm is applied to the resilience curve for $\lambda = 0.67$, resulting in the timepoints T_1 and T_2. The steady state in the resilience curve is presented in Figure 4 (b). The steady parts of the curve are shown in green, where the unsteady parts are shown in red. The detection of a steady state is successful, since 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.

Based on the resilience curve in Figure 4 (b), the resilience metrics are calculated. The degradation time, response time and recovery time measure 68, 95 and 60 minutes, respectively. The maximum impact measures 12.80 percentage points, and the performance loss measures approximately 1,848 minutes. The degradation profile measures 108.53 percentage points, 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 measures -47.44 percentage points, indicating a smaller, concave deviation from a linear recovery. This means performance recovered rapidly as many trains could be reinserted shortly after the restart was initiated.



(a) Resilience curve for different values of the performance weight

(b) Resilience curve with the identified steady state

Figure 4. Resilience curves for the example case, showing (a) the evolution of performance for different values of the performance weight, and (b) the identified steady state for the resilience curve with $\lambda = 0.67$.

4.3 Different types 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 as well. Inspection of over 100 randomly sampled disruptions (at least ten per cause) reveals that it is possible to distinguish between eight types of resilience curves, including the bathtub. The newly identified types are named the hammock shaped curve, plateau curve, steady state curve, gradual recovery curve, aftermath curve, timetable influenced curve and undefinable curve. Real examples of the different types of curves are presented in Figure 5.

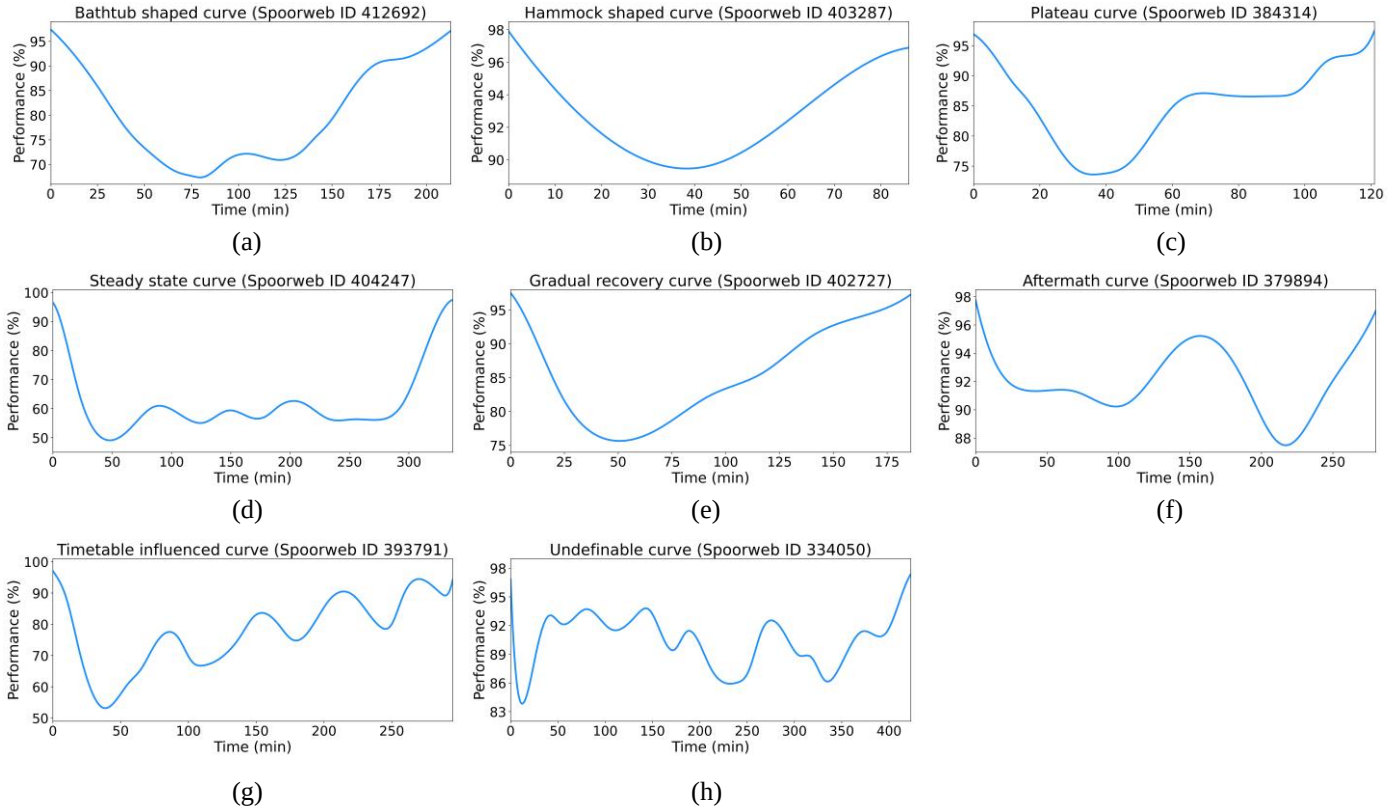


Figure 5. Examples of the different types of resilience curves.

Observations of the sampled disruptions suggest that certain types of resilience curves are more typical of one disruption cause than of another. The gradual recovery curve appears most typical of collisions, but is also occasionally observed for train defects. This seems to contradict the belief that collisions have a long and clear second phase due to the necessary clearing operations. The shape of the curve may be explained though by the fact that many trains are canceled shortly after a collision is reported. The number of cancellations could be unnecessarily high, creating a distorted version of the bathtub. The aftermath curve appears most typical of switch failures, but is occasionally observed for the other infrastructure related causes as well. The drop in performance in the end appears to be related to the permanent repair of the infrastructure, requiring a higher number of trains to be canceled temporarily to create a safe and accessible workspace for the mechanics. The undefinable curve appears most typical of section/signal failures, which may be explained by the sometimes unclear nature or location of the failure and the fact that this can result in a prognosis which is updated several times, creating uncertainty around the moment of restarting the train service.

The other types of resilience curves do not appear typical of a specific disruption cause, but are still worth discussing. The plateau curve could occur when part of the infrastructure becomes available again. The steady state curve could occur for particularly long disruptions where it is immediately clear that there will be little traffic for an extended period of time. The timetable influenced curve could occur for disruptions that involve relatively little traffic because they occur in a more isolated part of the network and/or occur in the early morning or late evening, when train frequencies are low, thus introducing an hourly pattern in the resilience curve.

4.4 Mean and median resilience curve per cause

To obtain a more general view of the resilience curve, the mean and median resilience curves are drawn for the studied disruption causes. Mean and median performance across a disruption cause 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. The mean and median curves and the central 80% range are shown in Figure 6 (a)-(e) for $\lambda = 0.67$, where the central 80% range is defined as the range of observations between the 10^{th} and the 90^{th} percentile. The mean curves are shown together in one plot in Figure 6 (f)-(h) for composite performance ($\lambda = 0.67$), punctuality ($\lambda = 0$) and traffic intensity ($\lambda = 1$), respectively.

The resilience curves in Figure 6 suggest that differences exist among disruptions of different causes. A preliminary conclusion would be that train defects are the least impactful single disruptions on average, whereas collisions are the most impactful single disruptions on average. This is most obvious in terms of traffic intensity, 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. Although the differences in terms of punctuality are smaller, it is observed that punctuality is affected more strongly on average for the infrastructure related causes than for train defects and collisions, particularly in the beginning. 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. The resulting delays can escalate quickly, especially on busy routes, which causes the lower punctuality.

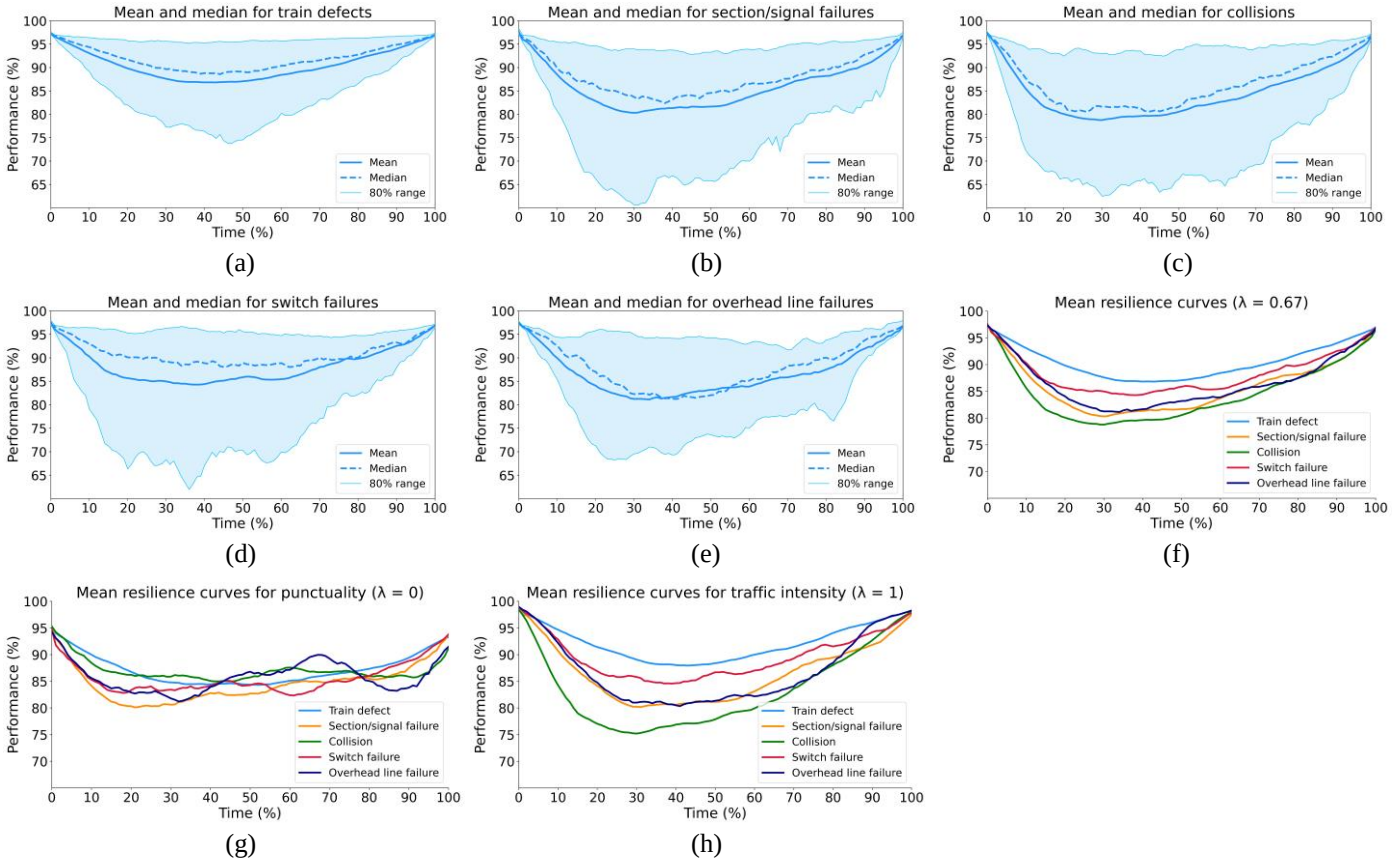


Figure 6. Mean and median resilience curve per disruption cause.

In addition, it is observed that the mean resilience curves do not necessarily resemble the shape of a bathtub, although the width of the 80% range shows that an arbitrary resilience curve could deviate significantly from the mean curve. The width of the 80% range is relatively small for train defects, which suggests train defects are the most consistent type of disruption. Another observation is that 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 behavior than in terms of recovery behavior. Also, the width of the 80% range for collisions, which is fairly constant from about $t = 20\%$ until $t = 60\%$, shows that certainly not all collisions have a gradual recovery curve, and that a significant number of collisions must have a resilience curve similar to the bathtub shaped curve or the steady state curve with a distinguishable second phase.

4.5 Comparison of resilience metrics across disruption causes

Group comparisons of the resilience metrics are performed to determine if differences indeed exist among disruptions of different causes. To get an overview of the values of the resilience metrics per cause, descriptive statistics are reported first. The mean, median and standard deviation (SD) of the resilience metrics are presented in Table 3. The table shows that a single disruption on average has a degradation time of 49.55 minutes, a response time of 79.17 minutes, a recovery time of 70.61 minutes, a maximum impact of 18.09 percentage points, a performance loss of 2,096 minutes, a degradation profile of -19.54 percentage points and a recovery profile of -54.46 percentage points. Note that the standard deviations are relatively large compared to the means. This suggests that the disruption dynamics are quite heterogeneous. Given the size of the standard deviations, the mean and median of the degradation and recovery profile are relatively close to zero, 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.

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

		DT (minutes)			RST (minutes)			RCT (minutes)			MI (percentage points)		
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

(Table 3 continued)

		PL (minutes)			DP (percentage points)			RP (percentage points)		
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

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 metric. Group comparisons in general are by definition two-sided tests, which is why 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.
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.

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

Metric	F-statistic	p-value	η-squared	H_0 rejected ($\alpha = 0.05$)
DT	4.770	1.77E-03	0.043	Yes
RST	22.352	1.06E-12	0.125	Yes
RCT	9.824	1.45E-06	0.081	Yes
MI	15.954	1.53E-09	0.129	Yes
PL	21.533	6.99E-12	0.169	Yes
DP	0.437	7.82E-01	0.012	No
RP	0.611	6.56E-01	0.008	No

In addition to Welch's ANOVA, the Games-Howell post hoc test is performed. 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 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 collisions have a significantly longer recovery time than train defects and overhead line failures. Train defects also have a significantly smaller maximum impact than collisions and section/signal failures. In terms of performance loss, collisions perform significantly worse than switch failures and train defects. Medium-sized differences involving overhead line failures 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 Figure 6, which already suggested that train defects are the least impactful single disruptions and collisions are the most impactful single 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 cancellations. For $\lambda = 1$, the degradation time and recovery time are shorter on average than for smaller values of λ, which means the transition phases (and the disruption as a whole) are found to last shorter if delays are not accounted for.

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

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 in terms of the response time, maximum impact and performance loss. On the contrary, the differences between groups for $\lambda = 1$ are less obvious in terms of the degradation time and recovery time. Note how there are some inconsistencies in the increasing or decreasing trend, for example with $\lambda = 0.80$. One explanation for these inconsistencies is that there may be a number of resilience curves which are sensitive to changes in λ because of 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 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

Based on the results of the case study, a general conclusion is that the disruption dynamics are quite heterogeneous. Even though differences exist among disruptions of different causes in terms of the resilience metrics, and some types of resilience curves are found to be more typical of one disruption cause than of another, there is still significant heterogeneity among disruptions within each group. Since composite performance represents the interaction between punctuality and traffic intensity, this heterogeneity also illustrates how complex the interaction between the two indicators can be. A common observation is 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 means delays can propagate less easily. If the train service is then restarted too fast or in a way that is infeasible for the TOCs, punctuality might drop again while traffic intensity already recovers. However, this is merely one of the possible scenarios. It could also be the case that punctuality is unaffected because trains are canceled immediately, or that punctuality and traffic intensity both recover at a similar rate, to give a few examples. Plotting the resilience curve for the separate indicators as well as for different combinations (i.e. different values of λ) creates a foundation to better understand this interaction.

Given that the resilience curve can have different shapes, even though each curve results from the same largely predefined and standardized process, one could ask whether a certain resilience curve behavior is preferred over another. Related to that, one could ask what the consequences are if a resilience curve does not resemble the bathtub shape like it is expected to. It should be noted that the bathtub model is mainly useful from a conceptual point of view. It helps translate the message that performance is temporarily lower than usual, but in practice, the real-time conditions in the network determine what is possible and what is not, 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 mainly creates predictability towards the TOCs and the passengers. When the constraints for the rescheduling of rolling

stock and crew in the second phase are clear, and TOCs can guarantee the availability of rolling stock and crew at the prognosed time of restart, then passengers know they can rely on the revisited timetable to reach their destination in spite of the disruption, which should be the end goal after all.

Regardless of the type of resilience curve, an array of factors might affect how 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 action, information supply or external conditions. An overview of explanatory factors per category is presented in Table 7. The number of factors alone illustrates how each disruption can be treated as a unique case. Still, the results of the resilience evaluation show that general conclusions can be drawn by studying all of these unique disruptions simultaneously, and that data-driven approaches do have potential to evaluate and improve resilience, even when part of the data are confounded by human interference.

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 action	Experience of the involved actors, proactive attitude of traffic controllers, change in workload, time pressure to reach the second phase, time pressure to restart the train service
Information supply	Swiftness of reporting the disruption, swiftness of communication throughout the chain, clarity about the cause and location, 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

6 Conclusion

In this paper, a data-driven resilience evaluation 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. The shape of each resilience curve is described by a set of resilience metrics, which are evaluated in group comparisons. By specifying the disruption cause as the grouping variable, differences and similarities can be found among disruptions of different 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 of the case study 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. Some resilience curves are fairly well behaved: they degrade, remain steady for some time and recover again, while other curves may show atypical behavior and can be quite unpredictable. With regard to the resilience metrics, significant differences were found among disruptions of different causes in terms of the degradation time, response time, recovery time, maximum impact and performance loss. The largest effect size was obtained for the performance loss. Post hoc tests showed that train defects are the least impactful single disruptions on multiple resilience metrics, whereas collisions are the most impactful single disruptions on multiple metrics. This finding is consistent with the mean resilience curves that were plotted for each disruption cause.

The resilience evaluation 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 connected to other disruptions. Furthermore, it was assumed that all delays and cancellations observed in the studied impact area were related to the disruption, while 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.

Future research could focus first of all on the interaction between punctuality and traffic intensity, which was described in terms of the composite performance indicator but is still not properly understood. Second, the same analysis could be repeated with newer data from after the pandemic, when traffic levels are back to normal again. Third, similar data-driven analyses could be performed for railway networks in other countries, where it would be particularly interesting to determine if anticipation-based or reaction-based disruption management should be preferred to build resilience. Fourth, it could be investigated how connected disruptions are best evaluated, and if different types are observed here as well. Fifth, it is worth investigating the expected effects of automatic train operation (ATO) on the disruption dynamics, since ATO removes much of the human element that is so clearly observable in the entire disruption management process.

Acknowledgements

The authors thank the colleagues in the traffic control department of ProRail for their support, openness and detailed discussions, with a special thanks to Wilco Tielman for supplying the necessary realization data.

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

Bešinović, N. (2020). Resilience in railway transport systems: a literature review and research agenda. *Transport Reviews, 40*(4), 457-478. https://doi.org/10.1080/01441647.2020.1728419

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. *Earthquake Spectra, 19*(4), 733-752. https://doi.org/10.1193%2F1.1623497

Büchel, B., Spanninger, T., & Corman, F. (2020). Empirical dynamics of railway delay propagation identified during the large-scale Rastatt disruption. *Scientific Reports, 10*, 18584. https://doi.org/10.1038/s41598-020-75538-z

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. *Transportation Research Part B: Methodological, 63*, 15-37. https://doi.org/10.1016/j.trb.2014.01.009

Cats, O., & Jenelius, E. (2014). Dynamic Vulnerability Analysis of Public Transport Networks: Mitigation Effects of Real-Time Information. *Networks and Spatial Economics, 14*, 435-463. https://doi.org/10.1007/s11067-014-9237-7

Chan, R., & Schofer, J. L. (2016). Measuring Transportation System Resilience: Response of Rail Transit to Weather Disruptions. *Natural Hazards Review, 17*(1), 05015004. https://doi.org/doi:10.1061/(ASCE)NH.1527-6996.0000200

Cimellaro, G. P., Reinhorn, A. M., & Bruneau, M. (2010). Seismic resilience of a hospital system. *Structure and Infrastructure Engineering, 6*(1-2), 127-144. https://doi.org/10.1080/15732470802663847

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Dalheim, Ø. Ø., & Steen, S. (2020). A computationally efficient method for identification of steady state in time series data from ship monitoring. *Journal of Ocean Engineering and Science, 5*(4), 333-345. https://doi.org/10.1016/j.joes.2020.01.003

Dekker, M. M., & Panja, D. (2021). Cascading dominates large-scale disruptions in transport over complex networks. *PLOS ONE, 16*(1), e0246077. https://doi.org/10.1371/journal.pone.0246077

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. *Public Transport, 9*, 343-364. https://doi.org/10.1007/s12469-017-0157-z

Gonçalves, L. A. P. J., & Ribeiro, P. J. G. (2020). Resilience of urban transportation systems. Concept, characteristics, and methods. *Journal of Transport Geography, 85*, 102727. https://doi.org/10.1016/j.jtrangeo.2020.102727

Goverde, R. M. P., & Hansen, I. A. (2013). Performance Indicators for Railway Timetables. *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings*, 301-306. https://doi.org/10.1109/ICIRT.2013.6696312

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. *European Journal of Epidemiology, 31*, 337-350. https://dx.doi.org/10.1007%2Fs10654-016-0149-3

Hosseini, S., Barker, K., & Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. *Reliability Engineering & System Safety, 145*, 47-61. https://doi.org/10.1016/j.ress.2015.08.006

Janić, M. (2015). Reprint of "Modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event". *Transportation Research Part A: Policy and Practice, 81*, 77-92. https://doi.org/10.1016/j.tra.2015.07.012

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. https://doi.org/10.1007/s11116-018-9875-6

Knoester, M. J. (2021). *A data-driven approach for evaluating the resilience of railway networks* [Master's thesis, Delft University of Technology].

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

Madni, A. M., Erwin, D., & Sievers, M. (2020). Constructing Models for System Resilience: Challenges, Concepts, and Formal Methods. *Systems, 8*(3). https://doi.org/10.3390/systems8010003

Malandri, C., Fonzone, A., & Cats, O. (2018). Recovery time and propagation effects of passenger transport disruptions. *Physica A: Statistical Mechanics and its Applications, 505*, 7-17. https://doi.org/10.1016/j.physa.2018.03.028

Mattsson, L.-S., & Jenelius, E. (2015). Vulnerability and resilience of transport systems – A discussion of recent research. *Transportation Research Part A: Policy and Practice, 81*, 16-34. https://doi.org/10.1016/j.tra.2015.06.002

Mertens, W., Pugliese, A., & Recker, J. (2017). *Quantitative Data Analysis: A Companion for Accounting and Information Systems Research*. Springer. https://doi.org/10.1007/978-3-319-42700-3

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. *International Journal of Production Research, 53*(22), 6736-6751. https://doi.org/10.1080/00207543.2015.1057296

Nicholson, G. L., Kirkwood, D., Roberts, C., & Schmid, F. (2015). Benchmarking and evaluation of railway operations performance. *Journal of Rail Transport Planning & Management, 5*(4), 274-293. https://doi.org/10.1016/j.jrtpm.2015.11.004

Ouyang, M., Dueñas-Osorio, L., & Min, X. (2012). A three-stage resilience analysis framework for urban infrastructure systems. *Structural Safety, 36-37*, 23-31. https://doi.org/10.1016/j.strusafe.2011.12.004

Parkinson, H. J., & Bamford, G. (2017). A journey into railway digitisation. *Stephenson Conference: Research for Railways 2017*, 333-340. https://www.researchgate.net/publication/320352839_A_journey_into_railway_digitisation

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. https://doi.org/10.3850/978-981-11-2724-3_0235-cd

Schipper, D., & Gerrits, L. (2018). Differences and similarities in European railway disruption management practices. *Journal of Rail Transport Planning & Management, 8*(1), 42-55. https://doi.org/10.1016/j.jrtpm.2017.12.003

Spiegler, V. L. M., Naim, M. M., & Wikner, J. (2012). A control engineering approach to the assessment of supply chain resilience. *International Journal of Production Research, 50*(21), 6162-6187. https://doi.org/10.1080/00207543.2012.710764

Uday, P., & Marais, K. (2015). Designing Resilient Systems-of-Systems: A Survey of Metrics, Methods, and Challenges. *Systems Engineering, 18*(5), 491-510. https://doi.org/10.1002/sys.21325

Wong, A., Tan, S., Chandramouleeswaran, K. R., & Tran, H. T. (2020). Data-driven analysis of resilience in airline networks. *Transportation Research Part E: Logistics and Transportation Review, 143*, 102068. https://doi.org/10.1016/j.tre.2020.102068

Woodburn, A. (2019). Rail network resilience and operational responsiveness during unplanned disruption: A rail freight case study. *Journal of Transport Geography, 77*, 59-69. https://doi.org/10.1016/j.jtrangeo.2019.04.006

Zhou, Y., Wang, J., & Yang, H. (2019). Resilience of Transportation Systems: Concepts and Comprehensive Review. *IEEE Transactions on Intelligent Transport Systems, 20*(12), 4262-4276. https://doi.org/10.1109/TITS.2018.2883766

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. *Transportation Research Record: Journal of the Transportation Research Board, 2599*(1), 70-80. https://doi.org/10.3141%2F2599-09

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. *Transportation Research Record: Journal of the Transportation Research Board, 2604*(1), 9-18. https://doi.org/10.3141%2F2604-02

Zilko, A. A., Kurowicka, D., & Goverde, R. M. P. (2016). Modeling railway disruption lengths with Copula Bayesian Networks. *Transportation Research Part C: Emerging Technologies, 68*, 350-368. https://doi.org/10.1016/j.trc.2016.04.018

Zobel, C. W. (2011). Representing perceived tradeoffs in defining disaster resilience. *Decision Support Systems, 50*(2), 394-403. https://doi.org/10.1016/j.dss.2010.10.001

B. Recommendations (in Dutch)

Naar aanleiding van de conclusies uit dit onderzoek zijn er aanbevelingen gedaan met betrekking tot de dataverwerking en bijsturingspraktijken binnen ProRail Verkeersleiding, welke hier worden besproken. Met betrekking tot de verwerking van realisatiedata en Spoorweb-data in Sherlock wordt het volgende aanbevolen:

1. **Zorg ervoor dat de meetmomenten altijd beschikbaar zijn** indien er een VSM is toegedeeld, zodat elke calamiteit op dezelfde manier kan worden geëvalueerd. Het gaat hierbij specifiek over de tijdstempels "eerste fase gereed" en "treindienst opgestart", die volgens de formele definitie respectievelijk het einde van de eerste en derde fase aangeven. Beide tijdstempels zijn afhankelijk van het feit of er al dan niet treinen zijn gekeerd, terwijl keren slechts één van de mogelijke bijsturingsacties is. Zonder keringen is er nog steeds een veerkrachtcurve te bepalen met een eerste, tweede en derde fase waarvoor de start- en eindtijd kunnen worden afgeleid, wellicht door middel van een definitiewijziging of zelfs door toepassing van de methodes die zijn gebruikt in deze thesis om de tijdspunten uit de curve zelf te halen.
2. **Beoordeel de start- en eindtijd van calamiteiten kritisch**, aangezien is gebleken dat calamiteiten gemiddeld genomen vaak eerder beginnen en later eindigen dan gerapporteerd. Met betrekking tot de starttijd wordt geadviseerd om de vijf minuten te herzien die standaard worden afgetrokken van het tijdstip wanneer een melding de meldkamer bereikt. Met betrekking tot de eindtijd wordt geadviseerd om verder te kijken dan de eerste gereden treinen bij het opstarten, zeker als er weinig treinseries bij betrokken zijn, aangezien het kan voorkomen dat vertragingen ontstaan tijdens het opstarten of dat het opstartplan alsnog niet maakbaar blijkt.
3. **Heroverweeg de definitie van einde eerste fase**, aangezien al voor aanvang van de experimenten bleek dat "eerste fase gereed" vaak geen nauwkeurige beschrijving geeft van het moment dat de veerkrachtcurve stabiliseert en de tweede fase is bereikt. In plaats daarvan zou het einde van de eerste fase simpelweg kunnen worden gedefinieerd als het moment waarop de eerste VSM is toegedeeld, wat uit de interviews naar voren kwam als de meer gangbare definitie. Bovendien bleek dit moment gemiddeld genomen redelijk nauwkeurig te zijn in vergelijking met het door het algoritme gevonden einde van de eerste fase, met name voor defect materieel.
4. **Verbeter de koppeling van klanthinder in Sherlock** tussen een Monitoring ID en een Spoorweb ID. Momenteel gaat deze koppeling in een aantal gevallen fout, bijvoorbeeld wanneer een eerdere verstoring (zonder VSM) heeft plaatsgevonden in hetzelfde gebied kort voor aanvang van de calamiteit waarop de klanthinder eigenlijk betrekking zou moeten hebben, of bijvoorbeeld wanneer de begrenzingspunten in het Monitoringdossier niet overeenkomen met die in het Spoorweb-dossier.
5. **Verbeter de VSM-beoordeling in Sherlock** door een extra check uit te voeren op de beoordeling die door de VLC in Spoorweb wordt ingevoerd. Momenteel is deze beoordeling niet altijd even betrouwbaar, en daarnaast komt het aantal beoordelingen niet altijd overeen met het aantal toegedeelde VSM's. De beoordeling zou moeten worden verbeterd wanneer Sherlock-data zouden worden gebruikt om een vergelijk te maken tussen calamiteiten waar een geschikte VSM beschikbaar was en calamiteiten waar deze niet beschikbaar was.
6. **Zorg ervoor dat realisatietijden en plantijden beschikbaar zijn** voor alle treinactiviteiten op elk dienstregelpunt, aangezien dit ten goede zou komen van de betrouwbaarheid van de prestatieberekening. Zonder realisatietijd kan niet worden bepaald of een activiteit punctueel was of niet, en zonder de meest recente plantijd kan niet met zekerheid worden gesteld dat alle treinpaden die gebruikt worden bij het bepalen van het impactgebied volledig kloppen. Hier wordt verder op ingegaan in Sectie 7.3, waarbij het gaat over de Planning in Tienden van Minuten (PINT) van NS-treinen.

Met betrekking tot de bijsturingspraktijken wordt het volgende aanbevolen:

1. **Creëer een stuk bewustwording** over het feit dat het snel bereiken van de volgende fase niet per se veel betekent, zowel voor de overgang van eerste naar tweede fase als voor de overgang van tweede naar derde fase. Meer specifiek betekent dit dat het snel doorlopen van de eerste fase niet per definitie leidt tot een betere (of slechtere) prestatie gedurende de rest van de calamiteit. Het zou dan ook de voorkeur verdienen om juist de tijd te nemen voor een weloverwogen en structureel maakbaar verkeersplan waarbij de kans op nieuwe vertragingen wordt beperkt, aangezien dit zou kunnen bijdragen aan een stabielere tweede fase en een kortere en meer vloeiende derde fase.
2. **Benadruk het belang van het snel delen van informatie** door de logistieke keten. Informatie omtrent een calamiteit hoort zo snel mogelijk te worden gedeeld, ongeacht hoe relevant de informatie is voor de actor die de informatie bezit, zodat andere actoren voor wie de informatie wel (of meer) relevant is hierop kunnen handelen. Het gaat hier voornamelijk om het begin van de versperring, waar het automatiseren van de melding richting de meldkamer door iets simpels als een druk op een knop kan helpen bij het sneller opstarten van de verschillende processen.
3. **Ontwerp vooraf gedefinieerde maatregelen voor de eerste fase** om de werkdruk voor verkeersleiders en treindienstleiders te verminderen. De standaardisatie van maatregelen voor de eerste fase zou verkeersleiders meer tijd kunnen geven om naar een optimale VSM toe te werken. Een dergelijke maatregel zou de eerste opheffingen en keringen kunnen specificeren op basis van de hypothetische locatie van een versperring. Aangezien timing cruciaal is bij het nemen van maatregelen in de eerste fase, zouden de acties per tijdvak van bijvoorbeeld vijf minuten in een basis-uur kunnen worden beschreven. Omdat de exacte toedracht van een versperring in het begin nog onduidelijk is, zou het verstandig zijn om een maatregel voor de eerste fase minstens zo beperkend te maken als een VSM voor een volledige baanvakstremming. De acties moeten daarnaast niet te veel conflicteren met de acties in een later toe te delen VSM, aangezien dit de overgang van eerste naar tweede fase zou bemoeilijken.
4. **Maak nauwkeurigere prognoses** van het moment waarop de infrastructuur wordt vrijgegeven en het opstarten van de treindienst kan beginnen. Een nauwkeurigere prognose is al een eerste stap richting een maakbaar opstartplan, aangezien dit de vervoerders voorziet van de nodige beperkingen om hun materieel- en personeelsplanning te optimaliseren tot het moment van opstarten. Indien de prognose wordt vervroegd, moet het bovendien zeker zijn dat de vervoerder zijn planning op orde heeft voordat er wordt begonnen met opstarten.
5. **Neem vaker een opstartkader op in een VSM** en betrek vervoerders hierbij. Indien er een opstartkader beschikbaar is, zou een verkeersleider meer tijd en vakmanschap kunnen steken in het voorbereiden van een structureel maakbaar opstartplan, waarbij de communicatielijnen kort en efficiënt worden gehouden. Een concept opstartkader kan nog steeds autonoom worden ontwikkeld op basis van de richtlijnen in het afwegingskader. Het zou echter verstandig zijn om deze versie daarna aan de vervoerders voor te leggen. Om de bruikbaarheid van het opstartkader te vergroten wordt tevens aanbevolen om het tijdsbestek vast te stellen waarin de opstart kan plaatsvinden, en daarbij te aan te geven welke series tegelijk kunnen worden opgestart en in welk tempo.
6. **Neem meer regie in de derde fase** om ervoor te zorgen dat het opstartplan wordt nageleefd en dat het soepel verloopt. Het klopt dat het beschikbaar stellen van de infrastructuur het voornaamste belang van ProRail is, waarna het aan de vervoerders is om hun treinen opnieuw te rijden. Desalniettemin is een goed uitgevoerd opstartplan met minimale vertragingen uiteindelijk in ieders voordeel, en creëert dit ook betere startcondities voor het geval dat er een volgende versperring in de buurt zou optreden.

7. **Vestig meer aandacht op het verbeteren van de afhandeling van aanrijdingen**, aangezien aanrijdingen met regelmaat voorkomen en de meest impactvolle opzichzelfstaande calamiteiten bleken te zijn. Het gaat hierbij niet enkel om aanrijdingen met een persoon, maar ook om andere soorten aanrijdingen, zeker aangezien het niet altijd gelijk duidelijk is of er een persoon bij betrokken is of niet. Aandacht zou uit kunnen gaan naar het versnellen van de melding bij een aanrijding en het kritisch beoordelen van het aantal treinen dat wordt opgeheven als noodmaatregel, ofwel het gebied waarin de treinen worden opgeheven. Als er geen (verdere) verbeteringen haalbaar zijn in het terugbrengen van de impact, zou aandacht moeten worden besteed aan manieren om aanrijdingen te voorkomen, zoals betere monitoring van personen op of nabij het spoor en het minder toegankelijk maken van de spoorbaan.
8. **Doe meer datagedreven onderzoek** naar de ontwikkeling van de systeemprestatie van het spoornetwerk en de kwaliteit van de logistieke afhandeling van calamiteiten, aangezien veerkracht steeds belangrijker zal worden naarmate het drukker wordt op het spoor. ProRail kan hierin bovendien meer initiatief nemen als probleemeigenaar. De benodigde data is beschikbaar, al dient deze eerst op een goede manier bij elkaar te worden gebracht en is er ruimte voor verbetering op het gebied van dataverwerking.
9. **Investeer in wiskundige optimalisatiemodellen** voor real-time bijsturingsmaatregelen. Net als de solver die wordt ontwikkeld voor het ontwerpen van VSM's, zouden modellen ook kunnen worden ontwikkeld voor maatregelen in de eerste en derde fase. Afzonderlijke modellen zouden de voorkeur krijgen boven één groot model om de overgang naar en van een VSM mogelijk te maken, aangezien de bijsturingsacties per fase verschillen, en om de bruikbaarheid ervan te waarborgen in het geval dat de rekentijd een beperkende factor is voor de toepassing van de modellen in de praktijk. Uiteindelijk zouden de modellen kunnen worden uitgebreid om ook vervangend vervoer mee te nemen in de maatregelen.

Het mag vermeld worden dat de aanbevelingen die hier zijn gepresenteerd grotendeels in lijn liggen met de operationalisering van de Koers van VL, die meest recentelijk in 2018 is beschreven als een toekomstvisie voor 2020. Hoewel de Koers van VL wordt gezien als een positieve ontwikkeling waarin veel bevindingen uit dit onderzoek zijn te herkennen, is het momenteel nog steeds een toekomstvisie. Als laatste aanbeveling wordt dan ook geadviseerd om de transitie naar een nog meer vooraf gedefinieerde en proactieve bijsturing daadwerkelijk te realiseren, zodat het niet slechts bij een visie blijft.

Behalve deze aanbevelingen zijn er ook mogelijkheden voor vervolgonderzoek binnen ProRail beschreven. Deze zijn te vinden in Sectie 7.3.

C. Overview of respondents

This table provides an overview in chronological order of the department and position of the respondents that were interviewed throughout the course of the research. In accordance with the General Data Protection Regulation, the respondents' names were not published.

#	Department	Position
1	PAB	Advisor performance analysis
2	VGB	Scenario maker
3	Staff VL	Staff worker
4	Staff VL	Advisor central staff
5	A&E	Manager analysis and advice
6	PAB	Advisor performance analysis
7	A&E	Project leader continuous improvement
8	A&E	Project leader continuous improvement
9	CMBO	National traffic controller
10	VL Post Midden-Nederland	Senior traffic controller
11	VL Post Zuid-Nederland	Regional traffic controller
12	A&E	Process leader continuous improvement
13	VGB	Scenario maker
14	CMBO	National traffic controller

Go back to page 5

D. Data collection summary

These tables provide detailed information about the collected data. This concerns traffic realization data, disruption log data, network data, VSM data and customer hindrance data.

Realization data

Description:	Traffic realization data
File name:	act_export_2018-12-09_2019-12-14
File size:	8.87 GB
Data format:	CSV
Retrieval date:	17-05-2021, 09:07
Retrieval method:	Request by e-mail
Contact person:	Wilco Tielman
Retrieved columns:	basic.drp, basic.drp_act, basic.drp_post, basic.plan, basic.treinnr, basic.treinnr_rijkarakter, basic.treinnr_vervoerder, basic.uitvoer, vklvos.plan_actueel, vklvos.plan_oorspronkelijk
Used columns:	All retrieved
Time period:	09-12-2018 until 14-12-2019

Disruption log data

Description:	Disruption log data for disruptions with a logistical record, retrieved from the Spoorweb viewer in Sherlock
File name:	export_spoorwebtabel_2019
File size:	1,650 kB
Data format:	Excel
Retrieval date:	19-04-2021, 19:14
Retrieval method:	Sherlock export
Sherlock version:	v2.46, 19-04-2021, 17:04
Retrieved columns:	All available
Used columns:	IncidentID, IncidentLabel, Dvlpost1, T_afsluit, T_voorval, T_gekozeneerstevsm, T_EindeIncidentICB, T_EindeIncident, T_opstartenmogelijk, T_Treindienstopgestart, Logistiek_VDBs, Logistiek_VDB_Begrenzingpunten
Time period:	09-12-2018 until 14-12-2019

Network data

Description:	Connection of each timetable point to adjacent timetable points
File name:	DONNA_71479_VER_1_IAUF_DRGLPT_VERBINDING
File size:	210 kB
Data format:	Text
Retrieval date:	04-05-2021, 10:53
Retrieval method:	Access through shared folder (VenD Datamarkt > 1_BU > 2019)

Go back to page 5

Go back to page 43

VSM data

Description:	Detailed information about the actions performed according to the VSM in each of the resilience phases
File name:	isvl-vsm-controle_[yyyy-mm-dd]
File size:	77 kB per file on average
Data format:	DAT
Retrieval date:	15-07-2021, 10:31
Retrieval method:	Request by e-mail
Contact person:	Wilco Tielman
Retrieved columns:	All available
Used columns:	Type, Treinserie_van, meldkaartnummer
Time period:	09-12-2018 until 14-12-2019

Customer hindrance data

Description:	Customer hindrance data measured in hindrance class and total minutes delay, retrieved from the VSM vs. customer hindrance analysis tool in Sherlock
File name:	export_klanthinder_2019
File size:	289 kB
Data format:	Excel
Retrieval date:	15-07-2021, 11:14
Retrieval method:	Sherlock export
Sherlock version:	v2.46, 13-07-2021, 12:21
Retrieved columns:	All available
Used columns:	Mon_ID, Klanthinder, Vtgm, SpoorWeb_ID
Time period:	09-12-2018 until 14-12-2019

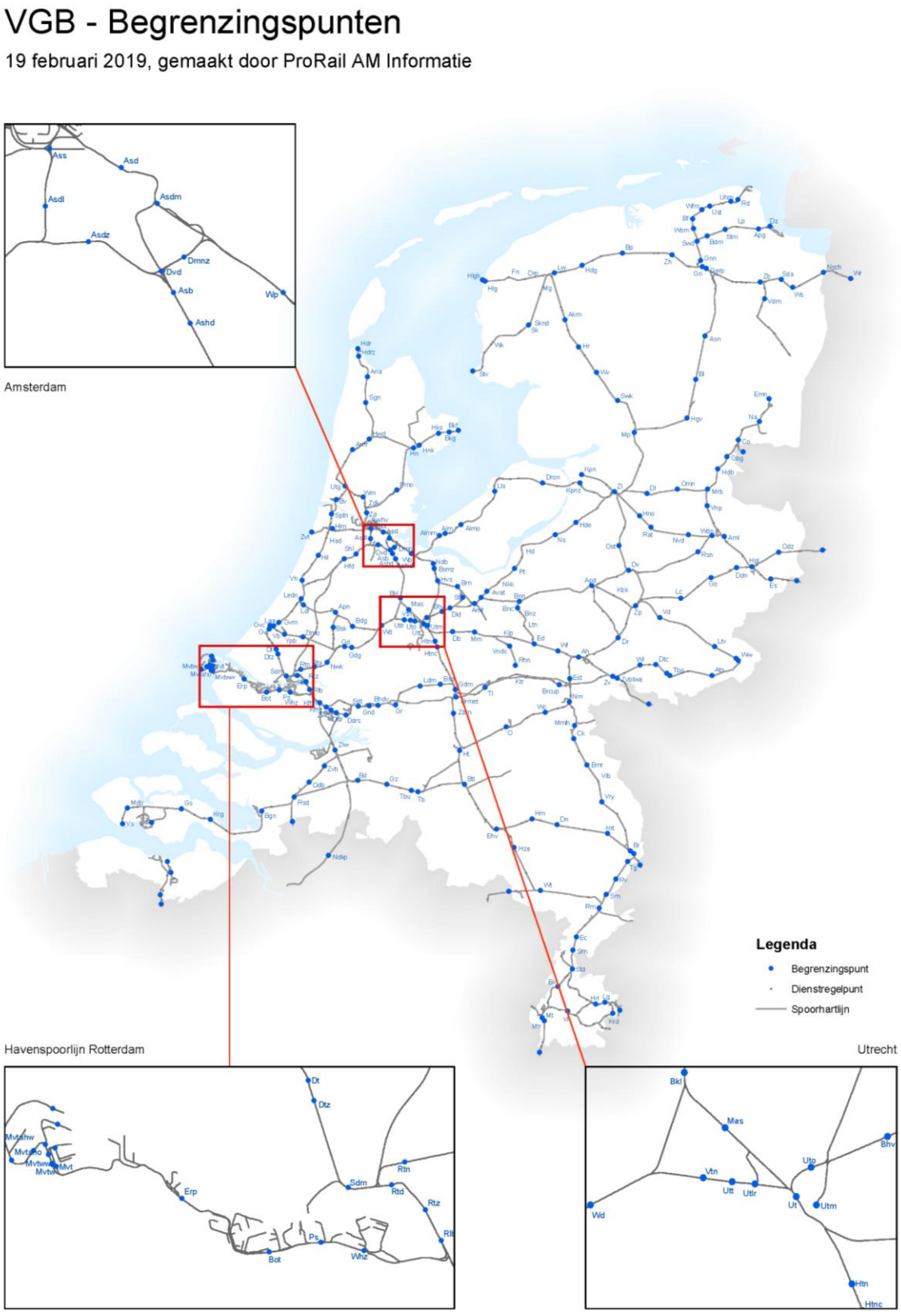
For each disruption, all relevant data other than the points on the resilience curve were stored in a dataframe with the following columns:

- Spoorweb_ID — Disruption ID in Spoorweb
- Date — Date when the disruption was reported
- Cause — Specific disruption cause
- Impact — Impact type as a result of the disruption
- Control_area — Traffic control area where the disruption occurred
- Boundaries — Boundary points according to the capacity reallocation
- N_boundaries — Number of unique boundary points
- Impact_area — Timetable points in the first and second impact area
- Area_size — Number of timetable points in the impact area
- Hindrance — Customer hindrance in total minutes delay
- VSM_available — Availability and adjustment of a VSM
- Series_in_VSM — Number of train series involved in the third phase
- Duration — Disruption length reported in Sherlock
- TT — Total time (observed disruption length)
- T0 — Observed start of disruption
- T0_S — Start of disruption reported in Sherlock
- T1 — Observed end of the first phase
- T1_S — End of the first phase reported in Sherlock
- T2 — Observed end of the second phase
- T2_S — End of the second phase reported in Sherlock
- T3 — Observed end of disruption
- T3_S — End of disruption reported in Sherlock
- DT — Degradation time
- RST — Response time
- RCT — Recovery time
- MI — Maximum impact
- PL — Performance loss
- DP — Degradation profile
- RP — Recovery profile

Go back to page 45

E. Map of boundary points

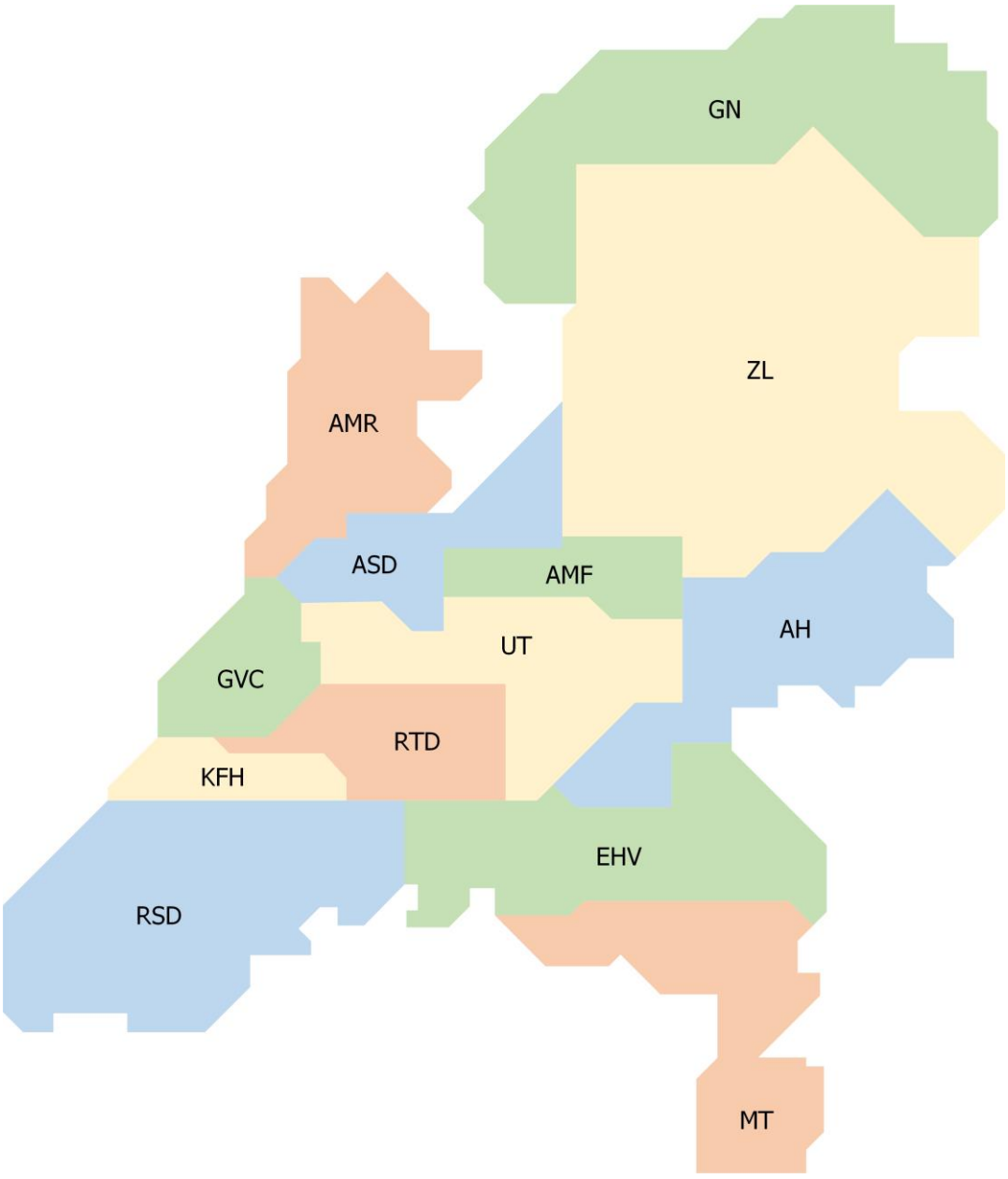
This map shows all timetable points in the Dutch railway network that can be identified as boundary points in case of a disruption.



Go back to page 25

F. Map of traffic control areas

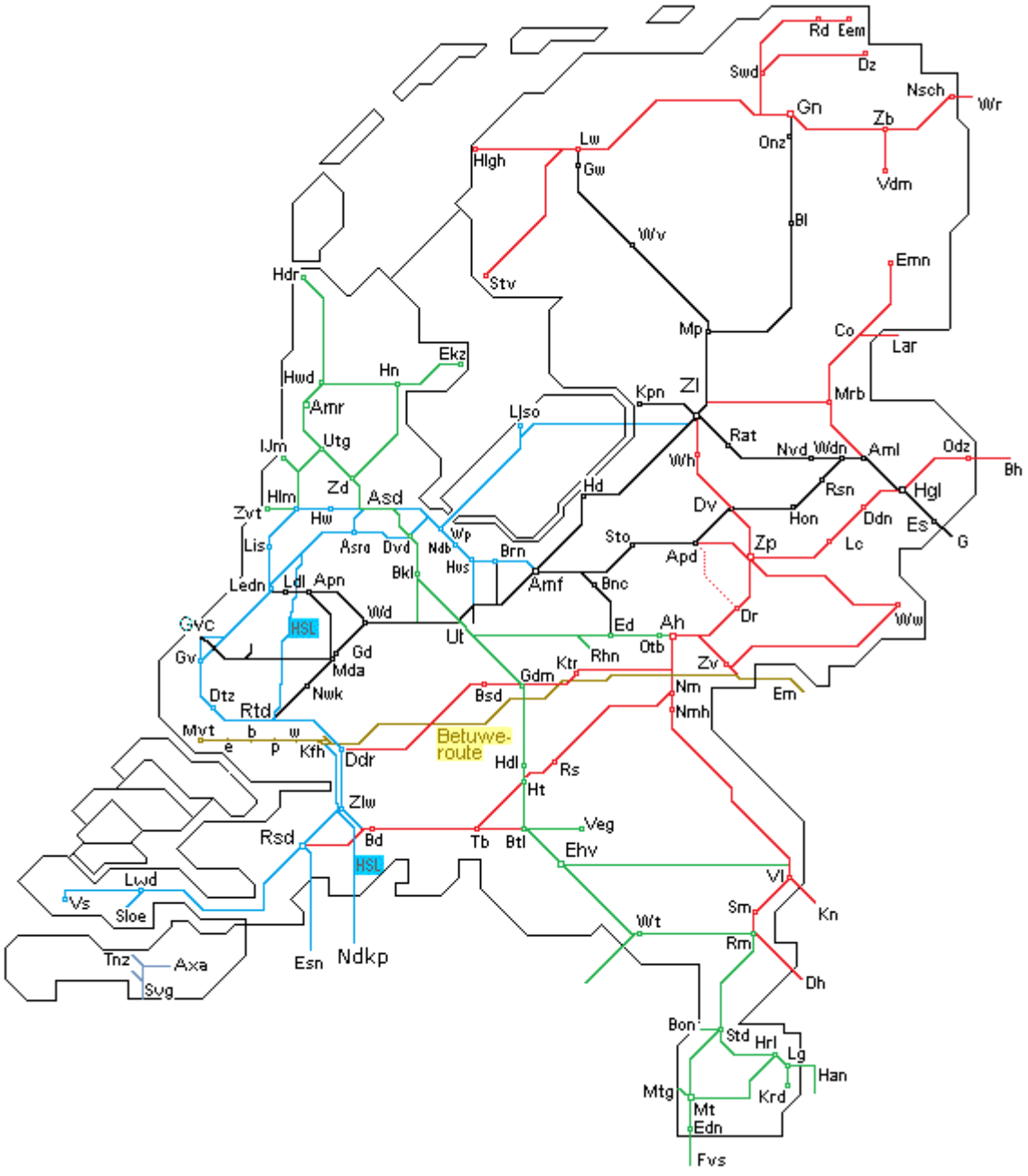
This map shows the control area of each of ProRail's regional traffic control centers. The area of Amersfoort (AMF) merged with Utrecht (UT) in 2019. Here it is shown separately since both areas appear in the analyzed datasets.



Go back to page 27

G. Map of passenger corridors

This map shows the division of the Dutch railway network into four main passenger corridors as used by the VGB team. The corridors are named *A2/A12* (green), *IJssellijn* (red), *Oudelijn* (blue) and *Groene Hart/Veluwe* (black).



Go back to page 30

H. Breadth first search algorithms

Below are the breadth first search algorithms that were developed to determine the first and second impact area for one, two and three start vertices, where:

- `DCs` is a list of decoupling points.
- `adjacency` is a dictionary that specifies the neighbors of each timetable point.
- `train_paths` is a string of all realized passenger train paths on a single day.
- `vertex` or `vertices` is/are the boundary point(s) from which to start.
- `impact` is a coded representation of the impact type.
- `max_area` defines the furthest impact area to include.

One start vertex

```
def bfs1(vertex, impact, max_area = 2, graph = adjacency):
    # initialize the algorithm
    queue  = [vertex]
    area   = {vertex: 1}
    parent = {vertex: None}

    # start the search
    while queue:
        v = queue.pop(0)
        for n in graph[v]:
            flag1 = False
            flag2 = False
            # search for neighbors of the start vertex
            if n not in area and v == vertex:
                if v in DCs and impact != 1:    # means: if not a full timetable point outage
                    area[n] = area[v] + 1
                else:
                    area[n] = area[v]
                parent[n] = v
                flag1 = True
            # search for neighbors of other vertices
            elif n not in area or (n in area and n in queue):
                if type(parent[v]) == str:
                    sequence1 = ','.join([parent[v], v, n])
                    sequence2 = ','.join([n, v, parent[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        flag2 = True
                elif type(parent[v]) == list:
                    sequence1 = ','.join([parent[v][0], v, n])
                    sequence2 = ','.join([n, v, parent[v][0]])
                    sequence3 = ','.join([parent[v][1], v, n])
                    sequence4 = ','.join([n, v, parent[v][1]])
                    if any(s in train_paths for s in [sequence1, sequence2, sequence3, sequence4]):
                        flag2 = True
            # mark neighbor as visited
            if flag2:
                if n not in area:
                    parent[n] = v
                else:
                    parent[n] = [parent[n], v]
                if v in DCs:
                    area[n] = area[v] + 1
                else:
                    area[n] = area[v]
                flag1 = True
            # add neighbor to the queue
            if flag1:
                if (area[n] == max_area and n in DCs) or n in queue:
                    pass
                else:
                    queue.append(n)
    return list(area.keys())
```

Two start vertices

```
def bfs2(vertices, max_area = 2, graph = adjacency):
    # initialize the algorithm
    queue   = [vertices[0]]
    visited = [vertices[0]]
    area    = {vertices[1]: 1}
    parent  = {vertices[0]: None}

    # PART 1: find the area between the start vertices
    # search for the second start vertex
    while vertices[1] not in visited:
        v = queue.pop(0)
        for n in graph[v]:
            flag = False
            # search for neighbors of the first start vertex
            if n not in visited and v == vertices[0]:
                visited.append(n)
                queue.append(n)
                parent[n] = v
            # search for neighbors of other vertices
            elif n not in visited or (n in visited and n in queue):
                if type(parent[v]) == str:
                    sequence1 = ','.join([parent[v], v, n])
                    sequence2 = ','.join([n, v, parent[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        flag = True
                elif type(parent[v]) == list:
                    sequence1 = ','.join([parent[v][0], v, n])
                    sequence2 = ','.join([n, v, parent[v][0]])
                    sequence3 = ','.join([parent[v][1], v, n])
                    sequence4 = ','.join([n, v, parent[v][1]])
                    if any(s in train_paths for s in [sequence1, sequence2, sequence3, sequence4]):
                        flag = True
            # mark neighbor as visited and add it to the queue
            if flag:
                if n not in visited:
                    visited.append(n)
                    parent[n] = v
                else:
                    parent[n] = [parent[n], v]
                if n not in queue:
                    queue.append(n)

    # trace back the traveled path
    while vertices[0] not in area:
        n = list(area.keys())[-1]
        if type(parent[n]) == str:
            area[parent[n]] = 1
        else:
            area[parent[n][0]] = 1
    # reinitialize the algorithm
    queue  = [vertices[0], vertices[1]]
    parent = {vertices[0]: list(area.keys())[-2], vertices[1]: list(area.keys())[1]}

    # PART 2: find the rest of the impact area
    # continue the search from the boundary points
    while queue:
        v = queue.pop(0)
        for n in graph[v]:
            flag = False
            # search for neighbors
            if n not in area or (n in area and n in queue):
                if type(parent[v]) == str:
                    sequence1 = ','.join([parent[v], v, n])
                    sequence2 = ','.join([n, v, parent[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        flag = True
```

```
                elif type(parent[v]) == list:
                    sequence1 = ','.join([parent[v][0], v, n])
                    sequence2 = ','.join([n, v, parent[v][0]])
                    sequence3 = ','.join([parent[v][1], v, n])
                    sequence4 = ','.join([n, v, parent[v][1]])
                    if any(s in train_paths for s in [sequence1, sequence2, sequence3, sequence4]):
                        flag = True
            # mark neighbor as visited and add it to the queue
            if flag:
                if n not in area:
                    parent[n] = v
                else:
                    parent[n] = [parent[n], v]
                if v in DCs:
                    area[n] = area[v] + 1
                else:
                    area[n] = area[v]
                if (area[n] == max_area and n in DCs) or n in queue:
                    pass
                else:
                    queue.append(n)
    return list(area.keys())
```

Three start vertices

```
import copy

def bfs3(vertices, max_area = 2, graph = adjacency):
    # initialize the algorithm
    queue1  , queue2  , queue3   = [vertices[0]]      , [vertices[0]]      , [vertices[1]]
    visited1, visited2, visited3 = [vertices[0]]      , [vertices[0]]      , [vertices[1]]
    area1   , area2   , area3    = {vertices[1]: 1}   , {vertices[2]: 1}   , {vertices[2]: 1}
    parent1 , parent2 , parent3  = {vertices[0]: None}, {vertices[0]: None}, {vertices[1]: None}

    # PART 1: find the area between the start vertices
    # search for the second start vertex starting from the first start vertex
    while vertices[1] not in visited1 and len(queue1) >= 1:
        v = queue1.pop(0)
        for n in graph[v]:
            if n not in visited1:
                if v == vertices[0]:
                    visited1.append(n)
                    queue1.append(n)
                    parent1[n] = v
                else:
                    sequence1 = ','.join([parent1[v], v, n])
                    sequence2 = ','.join([n, v, parent1[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        visited1.append(n)
                        queue1.append(n)
                        parent1[n] = v
    # search for the third start vertex starting from the first start vertex
    while vertices[2] not in visited2 and len(queue2) >= 1:
        v = queue2.pop(0)
        for n in graph[v]:
            if n not in visited2:
                if v == vertices[0]:
                    visited2.append(n)
                    queue2.append(n)
                    parent2[n] = v
                else:
                    sequence1 = ','.join([parent2[v], v, n])
                    sequence2 = ','.join([n, v, parent2[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        visited2.append(n)
                        queue2.append(n)
                        parent2[n] = v
    # search for the third start vertex starting from the second start vertex
    while vertices[2] not in visited3 and len(queue3) >= 1:
        v = queue3.pop(0)
        for n in graph[v]:
            if n not in visited3:
                if v == vertices[1]:
                    visited3.append(n)
                    queue3.append(n)
                    parent3[n] = v
                else:
                    sequence1 = ','.join([parent3[v], v, n])
                    sequence2 = ','.join([n, v, parent3[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        visited3.append(n)
                        queue3.append(n)
                        parent3[n] = v

    # trace back the traveled paths
    if vertices[1] in visited1:
        while vertices[0] not in area1:
            n = list(area1.keys())[-1]
            area1[parent1[n]] = 1
    if vertices[2] in visited2:
        while vertices[0] not in area2:
```

```
        n = list(area2.keys())[-1]
        area2[parent2[n]] = 1
if vertices[2] in visited3:
    while vertices[1] not in area3:
        n = list(area3.keys())[-1]
        area3[parent3[n]] = 1

# concatenate the traveled paths in both directions
area1_str = ','.join(list(area1.keys()))
area2_str = ','.join(list(area2.keys()))
area3_str = ','.join(list(area3.keys()))
area1_bw  = ','.join(list(reversed(list(area1.keys()))))
area2_bw  = ','.join(list(reversed(list(area2.keys()))))
area3_bw  = ','.join(list(reversed(list(area3.keys()))))
lengths   = [len(area1), len(area2), len(area3), len(area1_bw), len(area2_bw), len(area3_bw)]

# PART 2: check how the start vertices are connected
# for combinations where all paths are accurate
if all(x > 1 for x in lengths):
    flag01 = False
    flag02 = False
    flag12 = False
    if (area1_str in train_paths or area1_bw in train_paths) and (area2_str in train_paths or
        area2_bw in train_paths) and (area3_str in train_paths or area3_bw in train_paths):
        # check if one path contains the two other paths
        if area2_str in area1_str and area3_bw in area1_str:
            flag01 = True
        elif area1_str in area2_str and area3_str in area2_str:
            flag02 = True
        elif area1_bw in area3_str and area2_str in area3_str:
            flag12 = True
        # reinitialize the algorithm in case all paths are unique
        else:
            queue  = [vertices[0], vertices[1], vertices[2]]
            area   = {**area1, **area2, **area3}
            parent = {vertices[0]: [list(area1.keys())[-2], list(area2.keys())[-2]],
                      vertices[1]: [list(area1.keys())[1] , list(area3.keys())[-2]],
                      vertices[2]: [list(area2.keys())[1] , list(area3.keys())[1]]}
    # reinitialize the algorithm in case not all paths are unique
    else:
        if area1_str not in train_paths and area1_bw not in train_paths:
            if len(area2) > len(area3):
                flag02 = True
            else:
                flag12 = True
        elif area2_str not in train_paths and area2_bw not in train_paths:
            if len(area1) > len(area3):
                flag01 = True
            else:
                flag12 = True
        elif area3_str not in train_paths and area3_bw not in train_paths:
            if len(area1) > len(area2):
                flag01 = True
            else:
                flag02 = True
    if flag01:
        queue  = [vertices[0], vertices[1]]
        area   = copy.deepcopy(area1)
        parent = {vertices[0]: list(area.keys())[-2], vertices[1]: list(area.keys())[1]}
    if flag02:
        queue  = [vertices[0], vertices[2]]
        area   = copy.deepcopy(area2)
        parent = {vertices[0]: list(area.keys())[-2], vertices[2]: list(area.keys())[1]}
    if flag12:
        queue  = [vertices[1], vertices[2]]
        area   = copy.deepcopy(area3)
        parent = {vertices[1]: list(area.keys())[-2], vertices[2]: list(area.keys())[1]}
```

```
    # for combinations where not all paths are accurate
    # check which paths are accurate and reinitialize the algorithm
    else:
        queue = [vertices[0], vertices[1], vertices[2]]
        if (area1_str in train_paths or area1_bw in train_paths) and (area2_str in train_paths or
            area2_bw in train_paths) and all(x > 1 for x in [len(area1), len(area2)]):
            area   = {**area1, **area2}
            parent = {vertices[0]: [list(area1.keys())[-2], list(area2.keys())[-2]],
                      vertices[1]: list(area1.keys())[1],
                      vertices[2]: list(area2.keys())[1]}
        elif (area1_str in train_paths or area1_bw in train_paths) and (area3_str in train_paths or
              area3_bw in train_paths) and all(x > 1 for x in [len(area1), len(area3)]):
            area   = {**area1, **area3}
            parent = {vertices[0]: list(area1.keys())[-2],
                      vertices[1]: [list(area1.keys())[1], list(area3.keys())[-2]],
                      vertices[2]: list(area3.keys())[1]}
        elif (area2_str in train_paths or area2_bw in train_paths) and (area3_str in train_paths or
              area3_bw in train_paths) and all(x > 1 for x in [len(area2), len(area3)]):
            area   = {**area2, **area3}
            parent = {vertices[0]: list(area2.keys())[-2],
                      vertices[1]: list(area3.keys())[-2],
                      vertices[2]: [list(area2.keys())[1], list(area3.keys())[1]]}

    # PART 3: find the rest of the impact area (same as part 2 in bfs2)
    # continue the search from the boundary points
    while queue:
        v = queue.pop(0)
        for n in graph[v]:
            flag = False
            # search for neighbors
            if n not in area or (n in area and n in queue):
                if type(parent[v]) == str:
                    sequence1 = ','.join([parent[v], v, n])
                    sequence2 = ','.join([n, v, parent[v]])
                    if sequence1 in train_paths or sequence2 in train_paths:
                        flag = True
                elif type(parent[v]) == list:
                    sequence1 = ','.join([parent[v][0], v, n])
                    sequence2 = ','.join([n, v, parent[v][0]])
                    sequence3 = ','.join([parent[v][1], v, n])
                    sequence4 = ','.join([n, v, parent[v][1]])
                    if any(s in train_paths for s in [sequence1, sequence2, sequence3, sequence4]):
                        flag = True
            # mark neighbor as visited and add it to the queue
            if flag:
                if n not in area:
                    parent[n] = v
                else:
                    parent[n] = [parent[n], v]
                if v in DCs:
                    area[n] = area[v] + 1
                else:
                    area[n] = area[v]
                if (area[n] == max_area and n in DCs) or n in queue:
                    pass
                else:
                    queue.append(n)
    return list(area.keys())
```

Go back to page 45

I. Steady state detection algorithm

Below is the steady state detection algorithm that was developed to detect the steady state in the resilience curve, and thus, the start and end point of the second resilience phase. In the second line, it uses the column of the performance dataframe for the specified disruption ID.

```
import numpy as np
from scipy import stats

def ssd(ID, factor = 1.8, limit = 40, alpha = 0.01):
    # specify data
    data = performance[ID][performance[ID].notna()]
    X = list(data.index)
    Y = list(data)
    # define parameters
    n = int(factor * np.sqrt(len(X)))    # window length
    N = len(X)                           # number of measurements
    start_point = None
    search_limit = min(Y) + limit / 100 * (max(Y) - min(Y))

    # search for a start point
    for i in range(N):
        if Y[i] < search_limit:
            start_point = i
            break
    if start_point != None:
        # take an initial subset of the data
        end_of_window = start_point + n
        x = X[start_point: end_of_window]
        y = Y[start_point: end_of_window]
        state = np.zeros(len(X))
        # perform regression analysis on each consecutive window
        for i in range(start_point, N + 1 - n):
            model = stats.linregress(x, y)
            if model.pvalue > alpha and Y[i] < search_limit:
                state[i: end_of_window] = 1
            if i == N - n:
                break
            del x[0], y[0]
            x.append(X[end_of_window])
            y.append(Y[end_of_window])
            end_of_window += 1
        # generate output
        if any(s > 0 for s in state):
            T1 = next(i for i in X if state[i] == 1)
            T2 = next(i for i in reversed(X) if state[i] == 1)
            out = state, T1, T2
        else:
            out = None
    else:
        out = None
    return out
```

Go back to page 47

J. Overview of the data filtering steps

This table presents an overview of the data filtering steps that were applied to find the set of disruptions that could be analyzed. The table shows the number of remaining disruptions per step and the percentage of the data that is lost relative to the previous step.

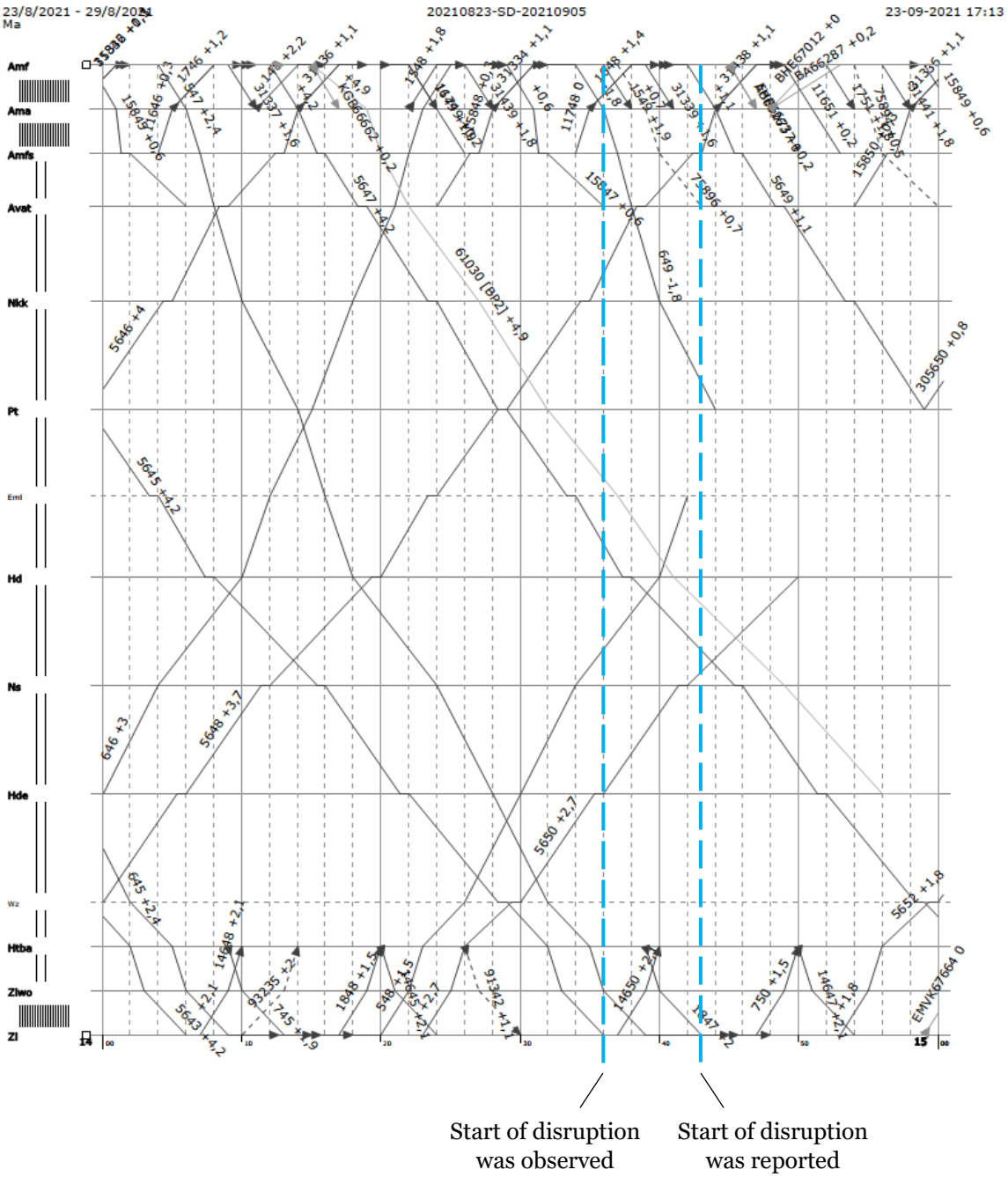
Filtering step →	1		2		3		4	
Excluded data →	▪ Disruptions on extreme days ▪ Disruptions in control area Kijfhoek		▪ Disruptions that do not match the top five causes		▪ Disruptions with a connection to one or more other disruptions		▪ Disruptions with an impact area size less than six ▪ Disruptions longer than ten hours ▪ Disruptions where "restart initiated" was not reported	
Specific cause	cases	% lost	cases	% lost	cases	% lost	cases	% lost
Train defect	742	-	742	0%	364	51%	346	5%
Section/signal failure	306	-	306	0%	156	49%	141	10%
Collision	275	-	275	0%	148	46%	146	1%
Switch failure	153	-	153	0%	50	67%	47	6%
Overhead line failure	65	-	65	0%	32	51%	26	19%
Total	1541[+]	-	1541	0%	750	51%	706	6%
Impact type	cases	% lost	cases	% lost	cases	% lost	cases	% lost
Full ttbp outage	55	-	23	58%	10	57%	10	0%
Partial ttbp outage	147	-	124	16%	48	61%	45	6%
Full line blockage	893	-	632	29%	344	46%	323	6%
Partial line blockage	805	-	703	13%	316	55%	299	5%
Reduced ttbp func.	32	-	23	28%	17	26%	17	0%
Reduced line func.	49	-	36	27%	15	58%	12	20%
Total	1981	-	1541	22%	750	51%	706	6%
Control area	cases	% lost	cases	% lost	cases	% lost	cases	% lost
AH	141	-	106	25%	61	42%	59	3%
AMF	94	-	76	19%	29	62%	29	0%
AMR	184	-	141	23%	90	36%	87	3%
ASD	190	-	147	23%	54	63%	52	4%
EHV	220	-	174	21%	91	48%	81	11%
GN	97	-	66	32%	55	17%	54	2%
GVC	102	-	80	22%	29	64%	28	3%
MT	112	-	83	26%	52	37%	51	2%
RSD	117	-	93	21%	44	53%	34	23%
RTD	269	-	234	13%	83	65%	81	2%
UT	234	-	170	27%	60	65%	59	2%
ZL	221	-	171	23%	102	40%	91	11%
Total	1981	-	1541	22%	750	51%	706	6%

[+] The causes of the remaining 440 disruptions are not shown in the first filtering step.

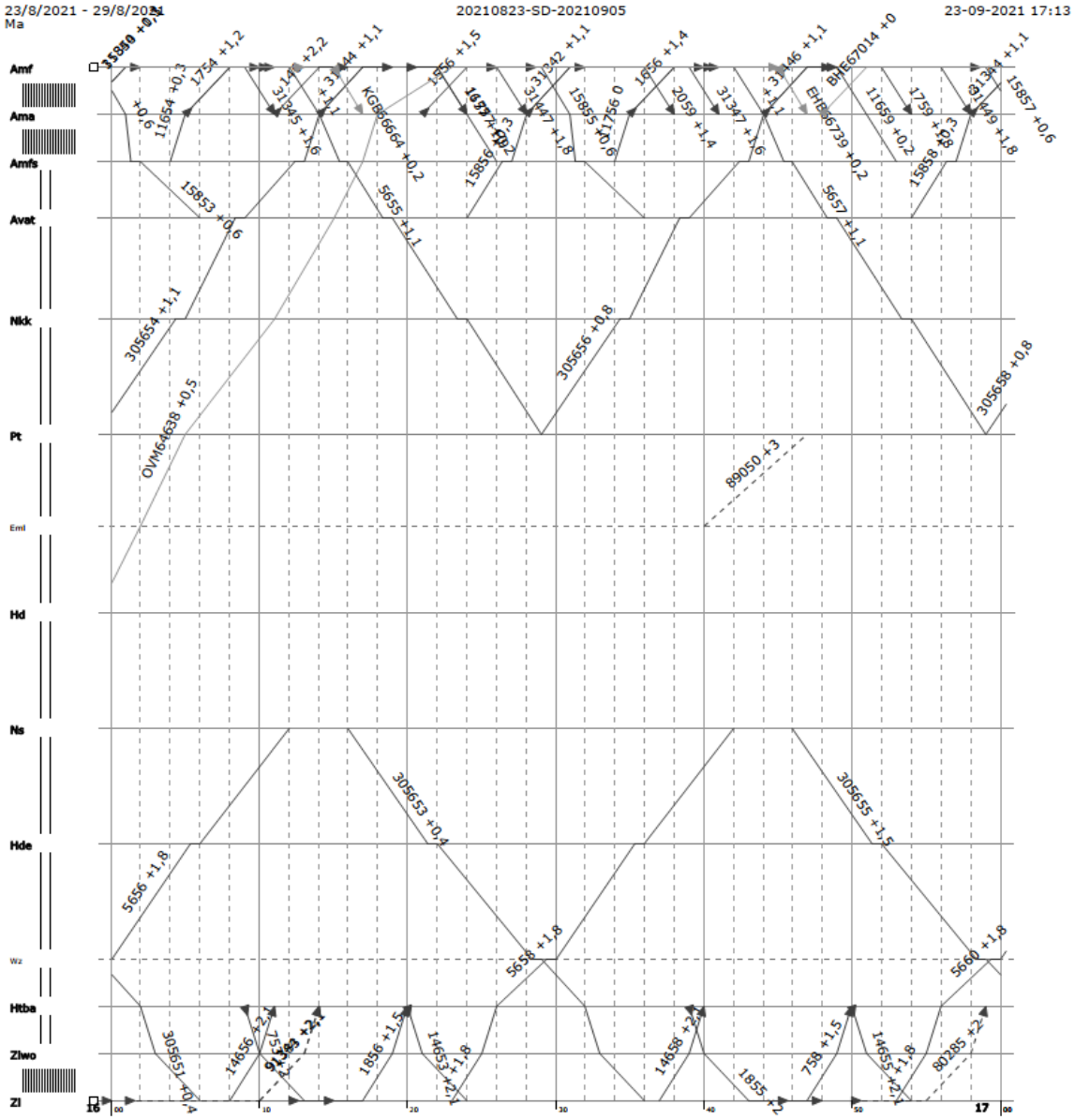
Go back to page 55

K. Time-distance diagrams for Amersfoort-Zwolle

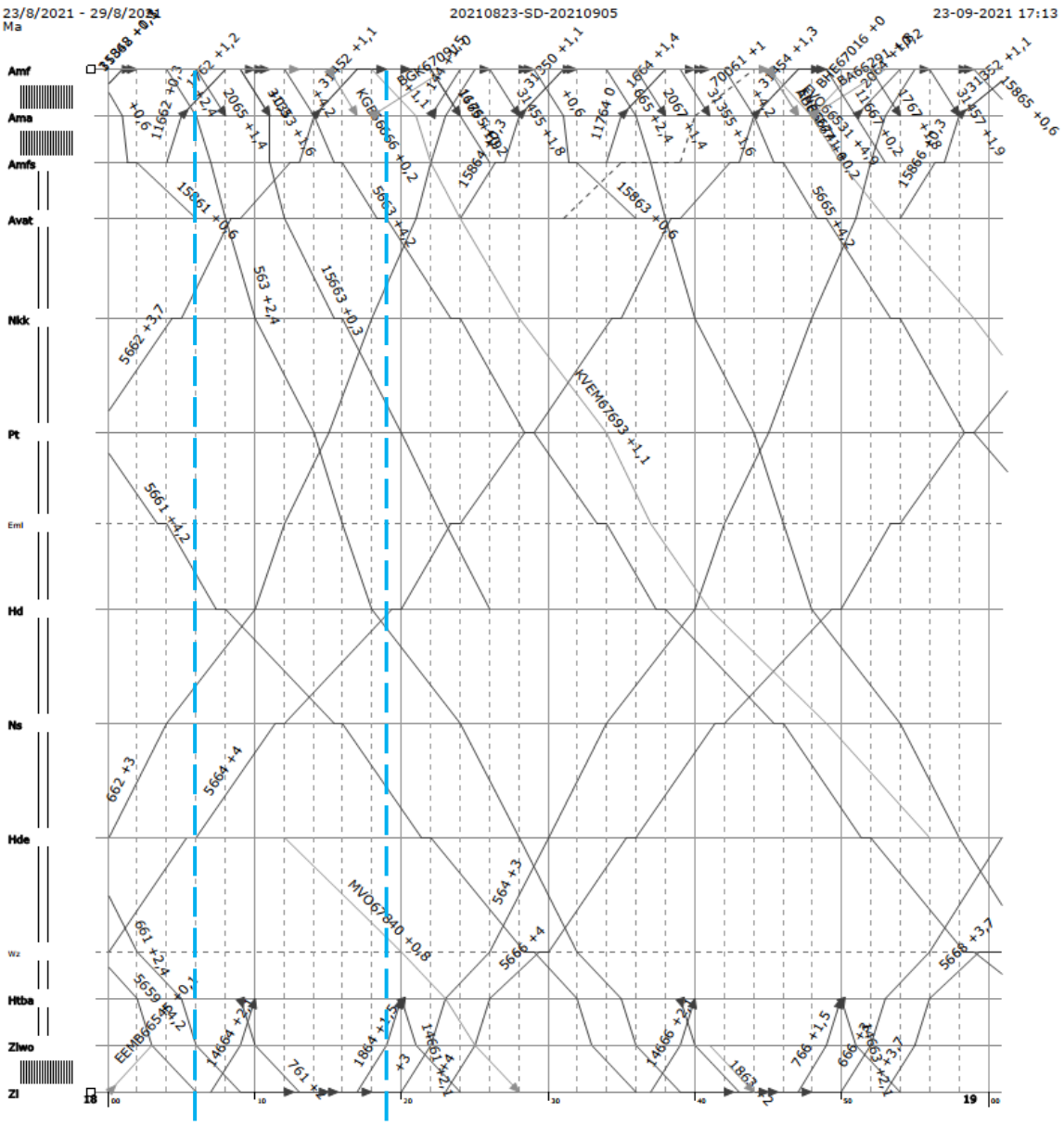
These time-distance diagrams present the realized train paths between Amersfoort and Zwolle for the disruption discussed in Section 5.2. The first diagram shows the time period in which the disruption occurred (14:00-15:00), along with the reported and observed start time.



The second diagram shows the time period in the middle of the disruption (16:00-17:00).



The third diagram shows the time period in which the disruption ended (18:00-19:00), along with the reported and observed end time.



SD Amf - Zl (Amersfoort-Zwolle) 5

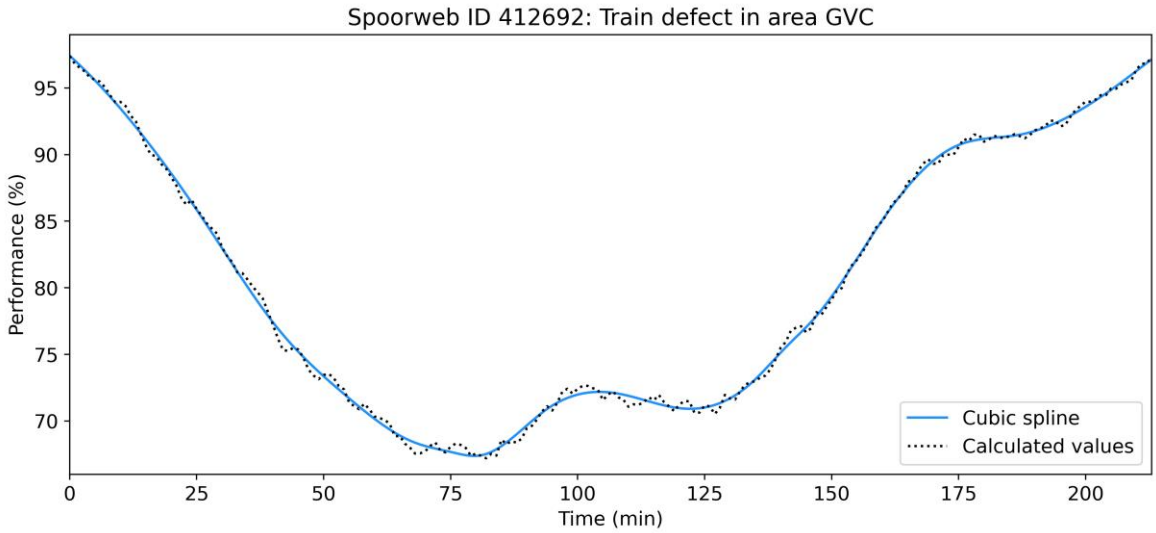
Go back to page 57

L. Examples of the types of resilience curves

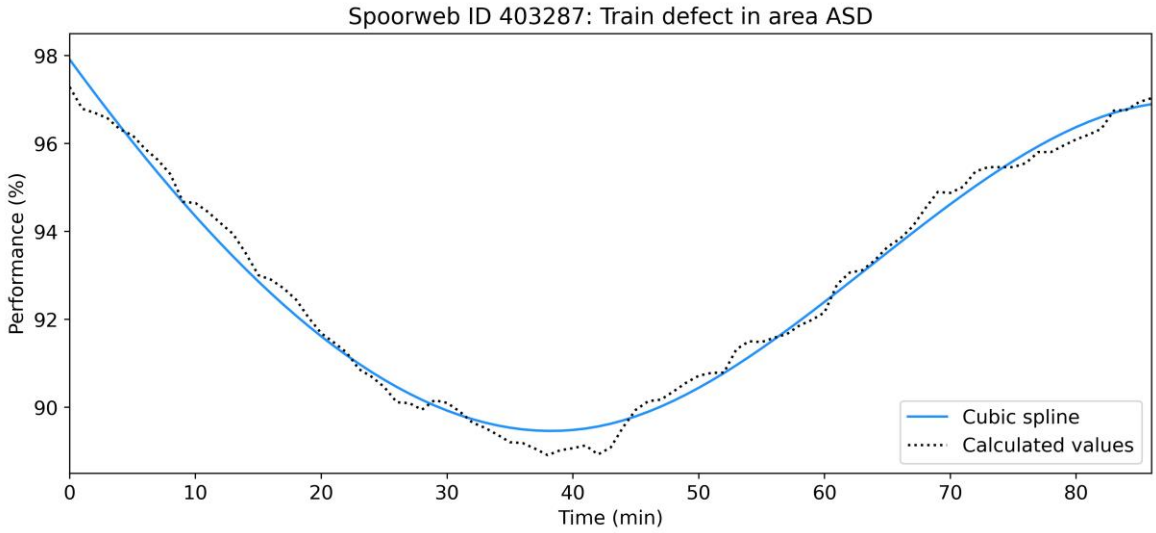
These figures illustrate the different types of resilience curves that were observed. Keep in mind that the axes are scaled differently per figure. A cubic spline, which is a concatenation of third degree polynomials, was fitted to obtain a smooth curve. The calculated points on the resilience curve are shown as a black, dotted line.

Go back to page 60

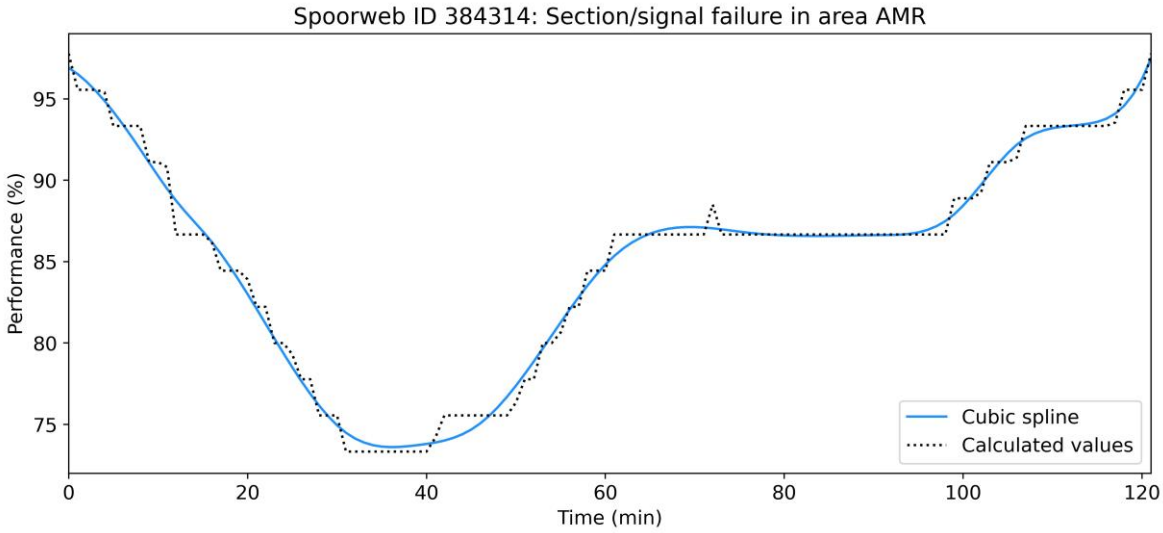
The bathtub shaped curve



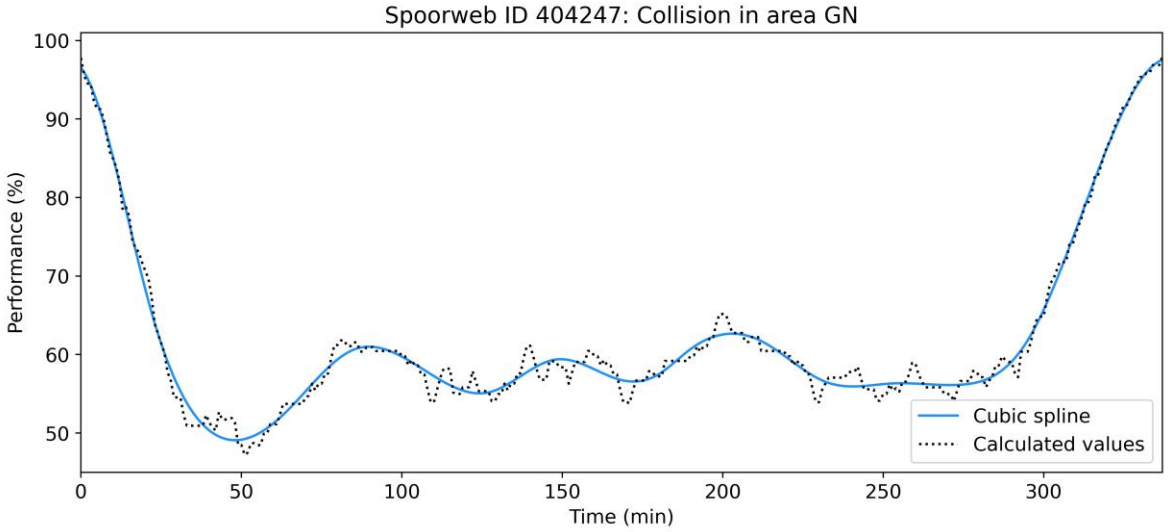
The hammock shaped curve



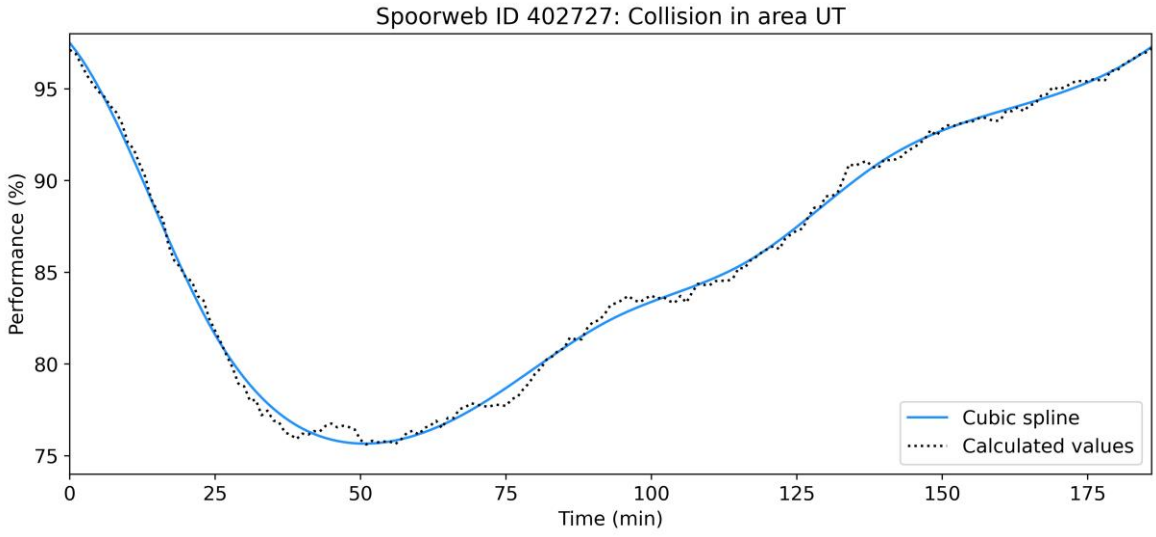
The plateau curve



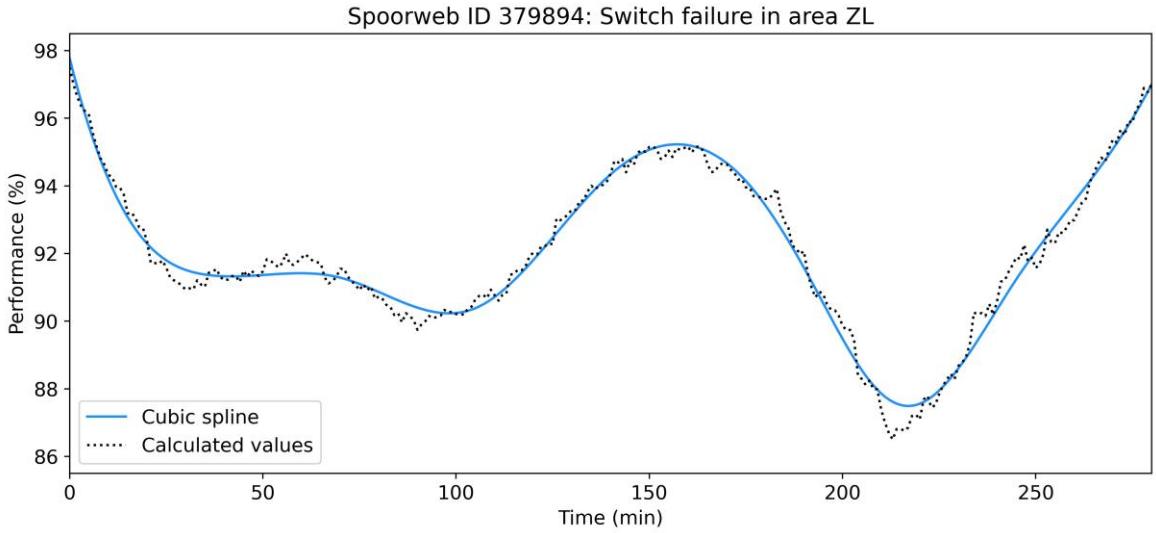
The steady state curve



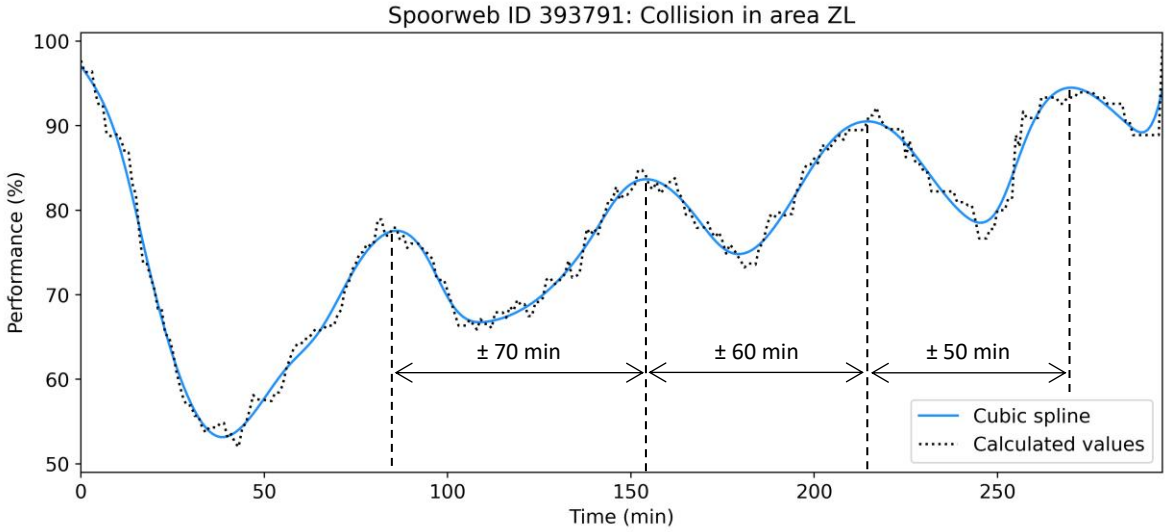
The gradual recovery curve



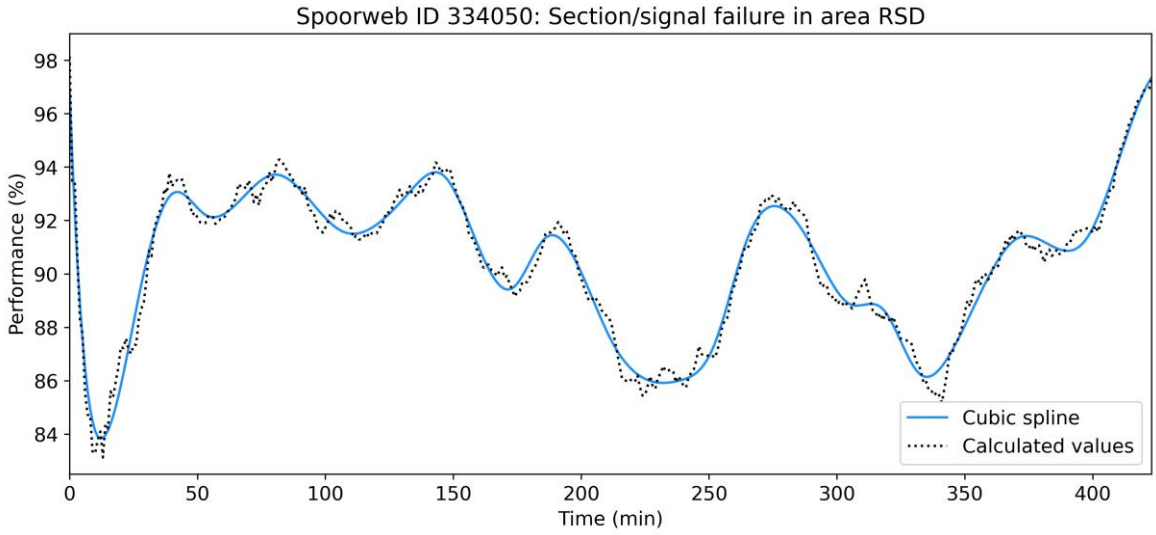
The aftermath curve



The timetable influenced curve

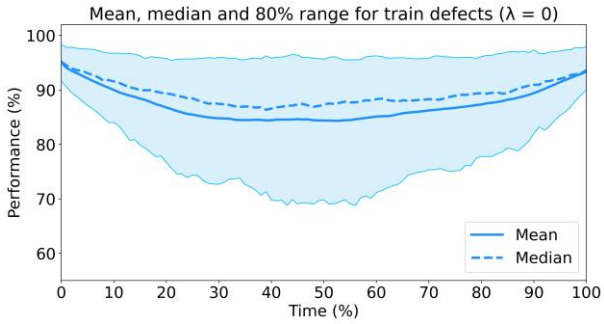


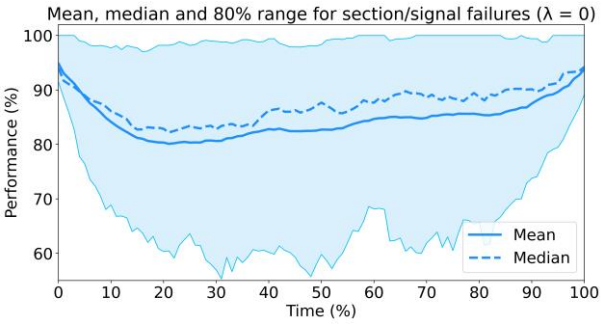
The undefinable curve

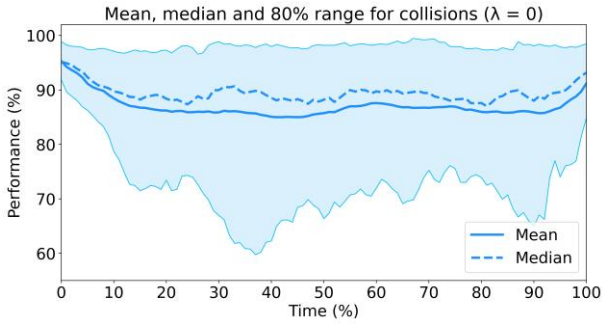


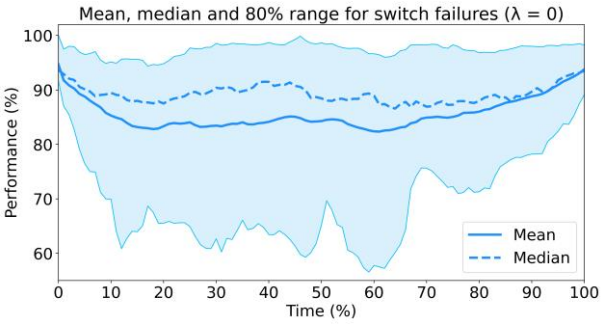
M. Mean resilience curves for separate indicators

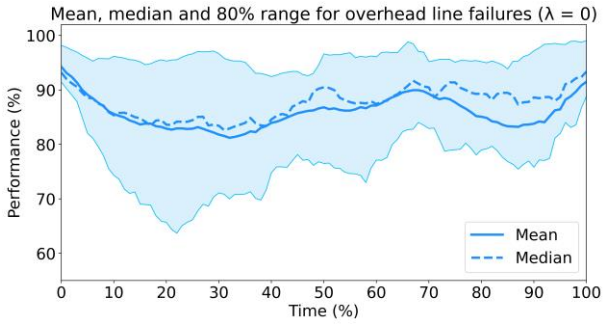
These figures present the mean and median resilience curve and the central 80% range for the two separate performance indicators. The first six figures show curves for punctuality ($\lambda = 0$). The last six figures show the curves for traffic intensity ($\lambda = 1$).

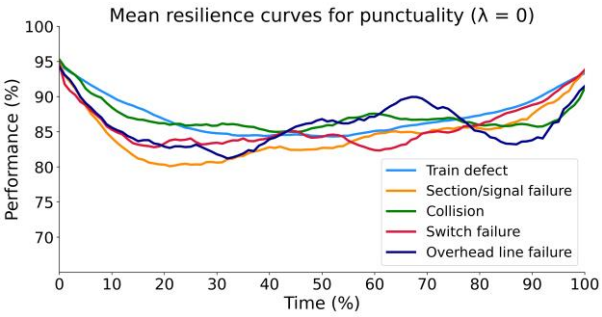


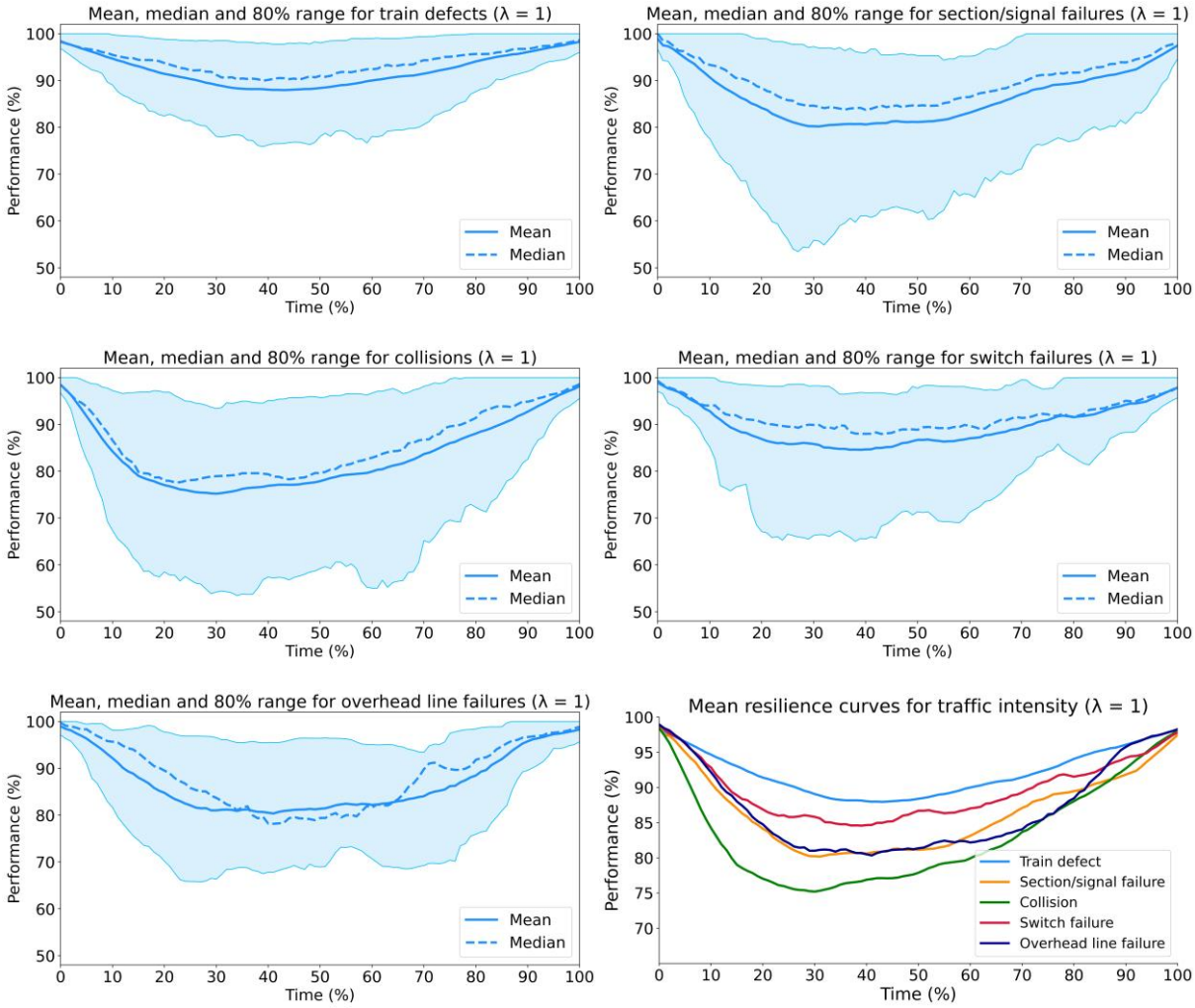












Go back to page 62

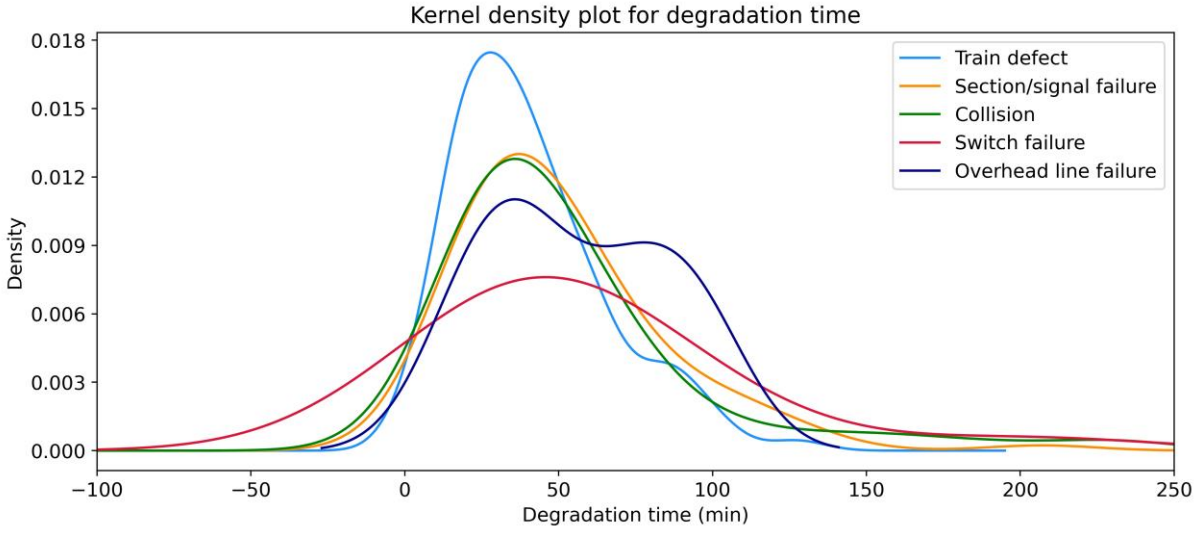
N. Assumptions check results

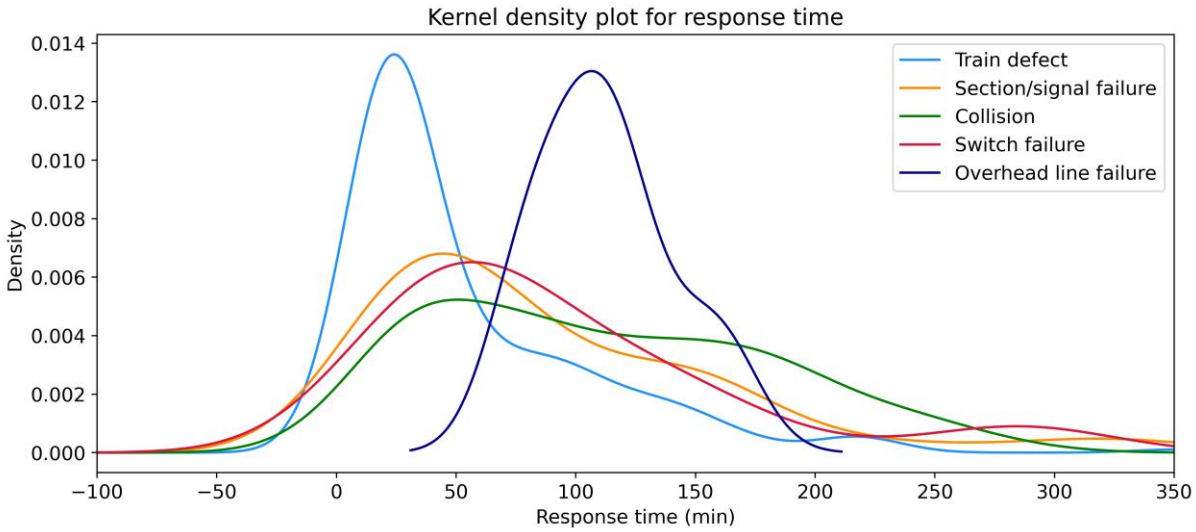
These tables and figures summarize the results of the assumptions check for Welch's ANOVA. The first table presents the Shapiro-Wilk test results regarding the assumption of normality. The second table presents the Levene test results regarding the assumption of equal variances. In addition, the kernel density plots for all metrics are presented.

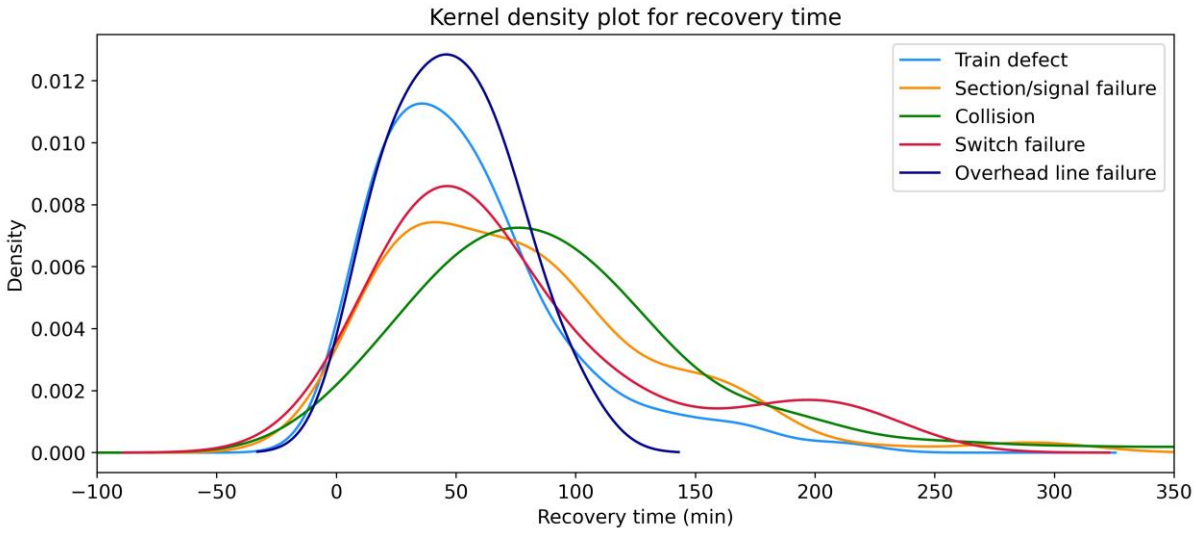
	DT		RST		RCT		MI	
Disruption cause	Shapiro W-statistic	Shapiro p-value	Shapiro W-statistic	Shapiro p-value	Shapiro W-statistic	Shapiro p-value	Shapiro W-statistic	Shapiro p-value
Train defect	0.937	1.06E-07	0.739	1.41E-17	0.884	2.30E-11	0.833	5.58E-14
Section/signal failure	0.689	4.85E-13	0.801	4.07E-10	0.887	4.83E-07	0.930	6.43E-05
Collision	0.698	9.32E-13	0.942	3.62E-04	0.889	6.79E-07	0.945	5.17E-04
Switch failure	0.541	3.71E-09	0.818	5.75E-05	0.815	4.98E-05	0.880	1.41E-03
Overhead line failure	0.921	1.75E-01	0.938	3.26E-01	0.973	8.85E-01	0.924	1.95E-01

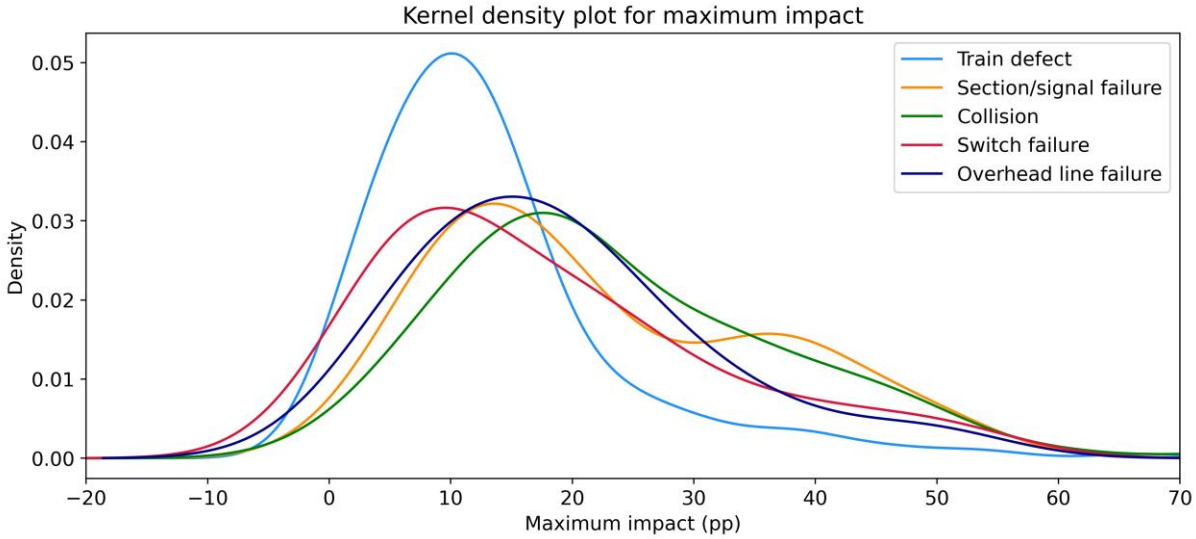
	PL		DP		RP	
Disruption cause	Shapiro W-statistic	Shapiro p-value	Shapiro W-statistic	Shapiro p-value	Shapiro W-statistic	Shapiro p-value
Train defect	0.606	2.83E-21	0.755	4.83E-17	0.685	3.16E-19
Section/signal failure	0.874	1.39E-07	0.393	3.61E-18	0.890	7.15E-07
Collision	0.911	6.89E-06	0.710	1.76E-12	0.783	1.42E-10
Switch failure	0.797	2.24E-05	0.297	1.28E-11	0.637	6.03E-08
Overhead line failure	0.774	1.25E-03	0.478	1.41E-06	0.764	9.54E-04

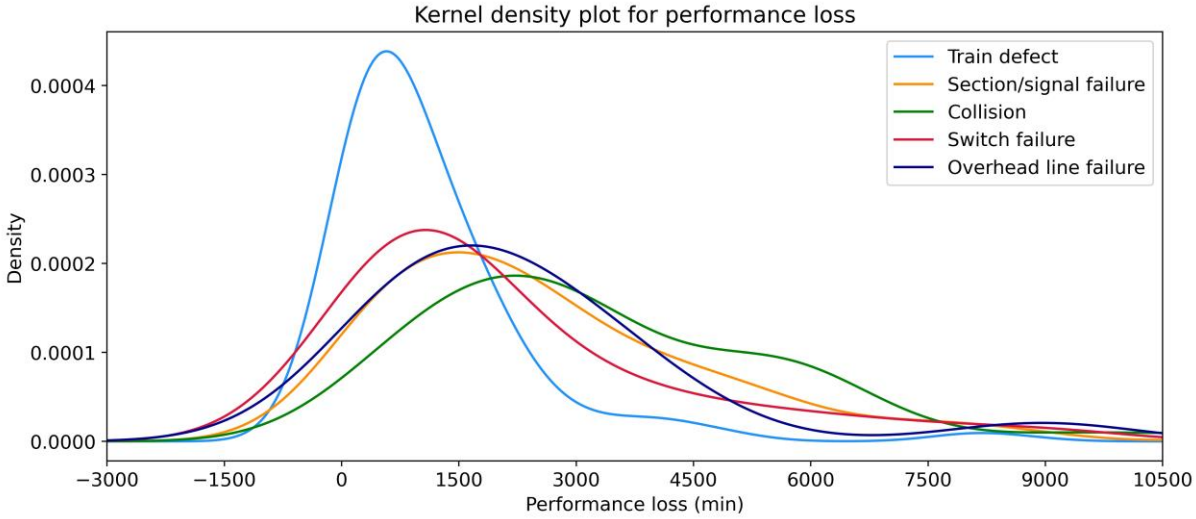
	DT	RST	RCT	MI	PL	DP	RP
Levene W-statistic	2.916	6.617	4.175	5.172	7.637	3.011	6.259
Levene p-value	2.11E-02	3.50E-05	2.50E-03	4.43E-04	6.00E-06	1.80E-02	6.60E-05

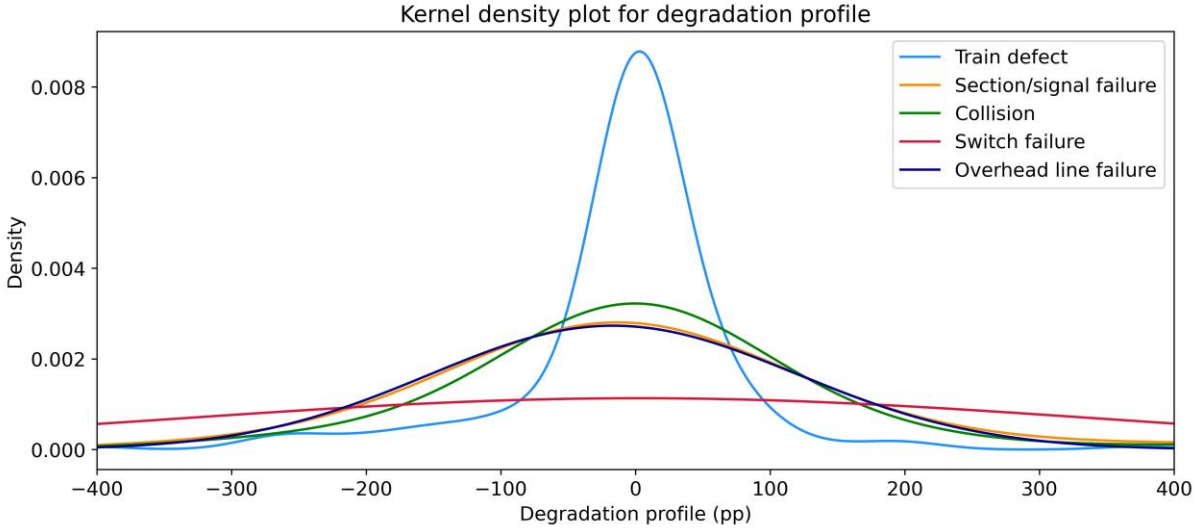


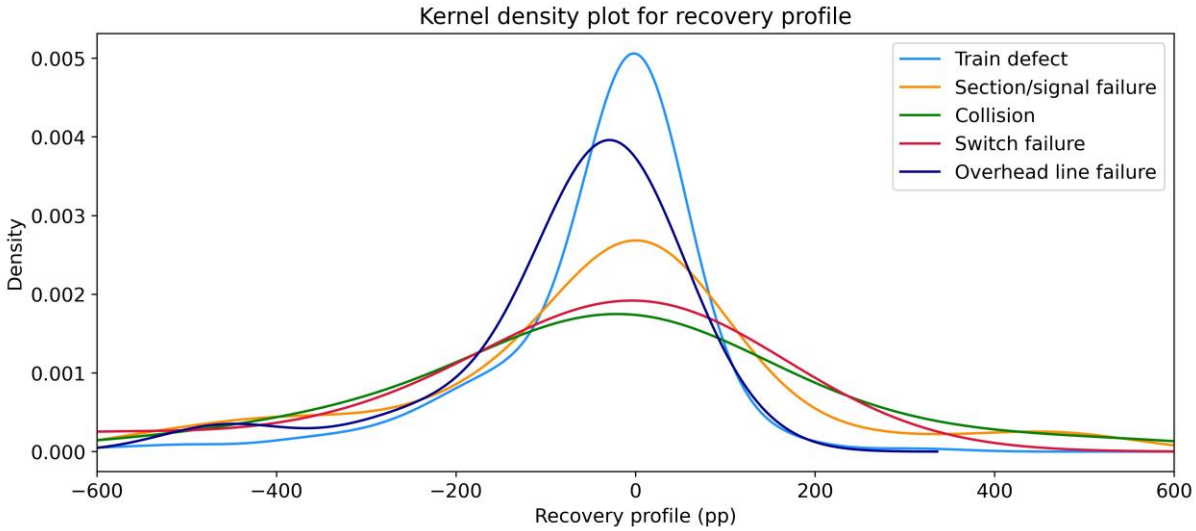

Kernel density plot for response time
Train defect
Section/signal failure
Collision
Switch failure
Overhead line failure
Density
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−100
−50
0
50
100
150
200
250
300
350
Response time (min)


Kernel density plot for recovery time
Train defect
Section/signal failure
Collision
Switch failure
Overhead line failure
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−100
−50
0
50
100
150
200
250
300
350
Recovery time (min)


Kernel density plot for maximum impact
Train defect
Section/signal failure
Collision
Switch failure
Overhead line failure
Density
0.05
0.04
0.03
0.02
0.01
0.00
−20
−10
0
10
20
30
40
50
60
70
Maximum impact (pp)







Go back to page 67

O. Post hoc test results

These tables present the full results of the Games Howell post hoc test for single disruptions and for single and connected disruptions. The pairwise comparisons for which Hedges' g was smaller than -0.5 or greater than 0.5 (indicating a medium to large effect) are printed in bold.

Metric:	DT								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	55.96	56.13	-0.17	8.87	-0.019	0.900	-0.005	0.499
Collision	Section/signal failure	55.96	53.61	2.35	6.82	0.344	0.900	0.049	0.514
Collision	Switch failure	55.96	69.88	-13.92	15.34	-0.908	0.888	-0.180	0.449
Collision	Train defect	55.96	40.62	15.34	5.41	2.838	0.042	0.351	0.598
Overhead line failure	Section/signal failure	56.13	53.61	2.52	8.54	0.295	0.900	0.079	0.522
Overhead line failure	Switch failure	56.13	69.88	-13.76	16.17	-0.851	0.900	-0.254	0.428
Overhead line failure	Train defect	56.13	40.62	15.51	7.46	2.079	0.274	**0.538**	0.649
Section/signal failure	Switch failure	53.61	69.88	-16.27	15.15	-1.074	0.796	-0.213	0.440
Section/signal failure	Train defect	53.61	40.62	12.99	4.85	2.677	0.063	0.330	0.592
Switch failure	Train defect	69.88	40.62	29.26	14.57	2.009	0.284	0.371	0.604
Metric:	RST								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	110.01	112.13	-2.11	9.91	-0.213	0.900	-0.057	0.484
Collision	Section/signal failure	110.01	91.04	18.97	10.75	1.765	0.398	0.253	0.571
Collision	Switch failure	110.01	92.79	17.22	14.50	1.188	0.733	0.236	0.567
Collision	Train defect	110.01	53.91	56.10	7.96	7.046	0.001	**0.871**	0.732
Overhead line failure	Section/signal failure	112.13	91.04	21.08	10.67	1.977	0.289	**0.530**	0.647
Overhead line failure	Switch failure	112.13	92.79	19.33	14.44	1.339	0.649	0.400	0.613
Overhead line failure	Train defect	112.13	53.91	58.22	7.86	7.410	0.001	**1.918**	0.913
Section/signal failure	Switch failure	91.04	92.79	-1.75	15.03	-0.117	0.900	-0.023	0.493
Section/signal failure	Train defect	91.04	53.91	37.14	8.89	4.176	0.001	**0.515**	0.642
Switch failure	Train defect	92.79	53.91	38.89	13.18	2.950	0.040	**0.545**	0.651
Metric:	RCT								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	93.33	48.25	45.08	9.08	4.964	0.001	**1.331**	0.828
Collision	Section/signal failure	93.33	78.60	14.74	8.74	1.686	0.447	0.242	0.568
Collision	Switch failure	93.33	76.24	17.10	12.00	1.425	0.600	0.283	0.580
Collision	Train defect	93.33	56.79	36.54	7.15	5.112	0.001	**0.632**	0.673
Overhead line failure	Section/signal failure	48.25	78.60	-30.35	8.60	-3.529	0.008	**-0.946**	0.250
Overhead line failure	Switch failure	48.25	76.24	-27.99	11.89	-2.353	0.146	**-0.702**	0.307
Overhead line failure	Train defect	48.25	56.79	-8.54	6.97	-1.225	0.715	-0.317	0.411
Section/signal failure	Switch failure	78.60	76.24	2.36	11.63	0.203	0.900	0.040	0.511
Section/signal failure	Train defect	78.60	56.79	21.81	6.52	3.343	0.009	0.412	0.615
Switch failure	Train defect	76.24	56.79	19.44	10.49	1.853	0.360	0.342	0.596
Metric:	MI								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	23.90	19.03	4.87	3.29	1.480	0.576	0.397	0.611
Collision	Section/signal failure	23.90	22.35	1.55	1.97	0.786	0.900	0.113	0.532

Collision	Switch failure	23.90	17.89	6.00	2.71	2.217	0.188	0.440	0.623
Collision	Train defect	23.90	13.24	10.66	1.60	6.663	0.001	**0.824**	0.720
Overhead line failure	Section/signal failure	19.03	22.35	-3.32	3.27	-1.015	0.828	-0.272	0.423
Overhead line failure	Switch failure	19.03	17.89	1.13	3.76	0.302	0.900	0.090	0.526
Overhead line failure	Train defect	19.03	13.24	5.79	3.06	1.896	0.357	0.491	0.636
Section/signal failure	Switch failure	22.35	17.89	4.45	2.68	1.660	0.468	0.329	0.592
Section/signal failure	Train defect	22.35	13.24	9.11	1.56	5.854	0.001	**0.721**	0.695
Switch failure	Train defect	17.89	13.24	4.66	2.42	1.924	0.322	0.356	0.600

Metric:	PL								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	3375.52	2358.22	1017.30	576.40	1.765	0.419	0.473	0.632
Collision	Section/signal failure	3375.52	2653.90	721.63	330.57	2.183	0.191	0.313	0.588
Collision	Switch failure	3375.52	2159.73	1215.80	446.27	2.724	0.062	**0.541**	0.650
Collision	Train defect	3375.52	1190.34	2185.18	266.96	8.185	0.001	**1.012**	0.763
Overhead line failure	Section/signal failure	2358.22	2653.90	-295.67	569.25	-0.519	0.900	-0.139	0.461
Overhead line failure	Switch failure	2358.22	2159.73	198.50	643.37	0.309	0.900	0.092	0.526
Overhead line failure	Train defect	2358.22	1190.34	1167.88	534.82	2.184	0.234	**0.565**	0.656
Section/signal failure	Switch failure	2653.90	2159.73	494.17	436.99	1.131	0.764	0.224	0.563
Section/signal failure	Train defect	2653.90	1190.34	1463.56	251.13	5.828	0.001	**0.718**	0.695
Switch failure	Train defect	2159.73	1190.34	969.39	391.09	2.479	0.117	0.458	0.627

Metric:	DP								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	-10.60	38.82	-49.43	59.85	-0.826	0.900	-0.221	0.437
Collision	Section/signal failure	-10.60	-24.66	14.06	36.43	0.386	0.900	0.055	0.516
Collision	Switch failure	-10.60	-109.25	98.64	114.75	0.860	0.900	0.171	0.548
Collision	Train defect	-10.60	-10.86	0.26	23.29	0.011	0.900	0.001	0.500
Overhead line failure	Section/signal failure	38.82	-24.66	63.49	62.44	1.017	0.827	0.273	0.577
Overhead line failure	Switch failure	38.82	-109.25	148.07	125.46	1.180	0.737	0.352	0.600
Overhead line failure	Train defect	38.82	-10.86	49.68	55.80	0.890	0.894	0.230	0.565
Section/signal failure	Switch failure	-24.66	-109.25	84.59	116.12	0.728	0.900	0.144	0.541
Section/signal failure	Train defect	-24.66	-10.86	-13.80	29.31	-0.471	0.900	-0.058	0.484
Switch failure	Train defect	-109.25	-10.86	-98.39	112.69	-0.873	0.900	-0.161	0.454

Metric:	RP								
Group A	Group B	Mean(A)	Mean(B)	(A – B)	SE	t-value	p-value	Hedges' g	CLES
Collision	Overhead line failure	-41.10	-74.74	33.64	48.56	0.693	0.900	0.186	0.553
Collision	Section/signal failure	-41.10	-43.07	1.97	43.12	0.046	0.900	0.007	0.502
Collision	Switch failure	-41.10	-124.36	83.25	64.98	1.281	0.680	0.254	0.572
Collision	Train defect	-41.10	-52.90	11.80	38.14	0.309	0.900	0.038	0.511
Overhead line failure	Section/signal failure	-74.74	-43.07	-31.67	39.12	-0.810	0.900	-0.217	0.439
Overhead line failure	Switch failure	-74.74	-124.36	49.62	62.40	0.795	0.900	0.237	0.568
Overhead line failure	Train defect	-74.74	-52.90	-21.83	33.54	-0.651	0.900	-0.168	0.452
Section/signal failure	Switch failure	-43.07	-124.36	81.29	58.27	1.395	0.618	0.276	0.578
Section/signal failure	Train defect	-43.07	-52.90	9.84	25.04	0.393	0.900	0.048	0.514
Switch failure	Train defect	-124.36	-52.90	-71.45	54.68	-1.307	0.668	-0.241	0.432

Go back to page 68