Robustness indicators and capacity models for railway networks

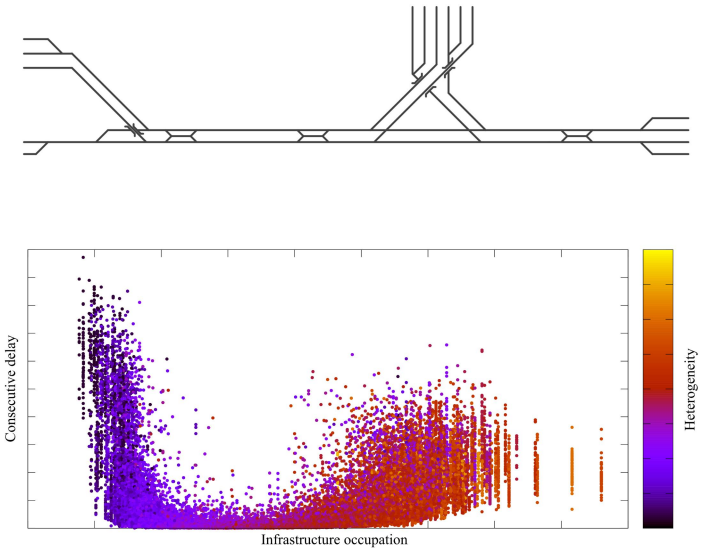
Jensen, Lars Wittrup; Nielsen, Otto Anker; Landex, Alex

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Robustness indicators and capacity models for railway networks

PhD Thesis



Lars Wittrup Jensen

2015

DTU Transport
Department of Transport

Robustness indicators and capacity models for railway networks

PhD thesis by

Lars Wittrup Jensen
Department of Transport
Technical University of Denmark

Supervisors:

Professor Otto Anker Nielsen
Department of Transport
Technical University of Denmark

Chief Consultant & PhD Alex Landex
Planning & Rolling Stock
Ramboll Denmark A/S

Kongens Lyngby

December 2015

Preface

This thesis is the result of the PhD project "Robustness indicators and capacity models for railway networks" carried out at the Department of Transport, Technical University of Denmark in the period 2012-2015. The thesis is submitted to fulfill the requirements for the PhD degree in engineering science at the Technical University of Denmark. The PhD project has been supervised by Professor Otto Anker Nielsen, and co-supervised by Chief Consultant & PhD Alex Landex.

The PhD project is part of the RobustRailS research project under work package 3 which is entitled "Robustness in the rail operational process". The RobustRailS project seeks to develop quantitative methods to measure robustness and use these methods to improve robustness of railways to attract more passengers to rail transport.

During the PhD project, the following articles have been published, submitted, or presented at conferences:

1. L. W. Jensen and A. Landex. Measuring Robustness of Timetables at Stations using a Probability Distribution. In *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis*, Copenhagen, 2013a

2. *Where parts have been included in the article:*
   A. Landex and L. W. Jensen. Measures for track complexity and robustness of operation at stations. *Journal of Rail Transport Planning & Management*, 3(1): 22–35, 2013

3. L. W. Jensen, A. Landex, and O. A. Nielsen. Evaluation of Robustness Indicators using Railway Operation Simulation. In C. Brebbia, N. Tomii, P. Tzieropoulos, and J. Mera, editors, *Computers in Railways XIV*, pages 329–339, Rome, 2014. WITpress. doi: 10.2495/CR140271

4. L. W. Jensen, A. Landex, O. A. Nielsen, L. G. Kroon, and M. Schmidt. Strategic assessment of capacity consumption in railway networks: Framework and model. *Submitted to: Transportation Research: Part C*, 2016

5. *Which is an extension of the article*:
   L. W. Jensen, A. Landex, and O. A. Nielsen. Assessment of Stochastic Capacity Consumption in Railway Networks. In *Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis*, Tokyo, 2015

6. L. W. Jensen. An optimisation framework for determination of capacity in railway networks. In *Proceedings of Conference on Advanced Systems in Public Transport (CASPT)*, Rotterdam, 2015

The six articles listed above form the backbone of this PhD thesis. In addition to these articles, unpublished extensions and results have also been included in the thesis. Furthermore, the articles of Jensen (2013) and Jensen and Landex (2013b) have been presented at the Annual Transport Conference at Aalborg University 2013, which are not directly related to the topic of this PhD thesis, and are thus not included.

Lars Wittrup Jensen
Kongens Lyngby, December 2015

Acknowledgements

"Though the course may change sometimes, rivers always reach the sea"

– Led Zeppelin (1975)

These past three years has been a great journey of fun, frustration, uncertainty, new ideas, and exciting results that is now coming to an end. Obviously, I would not have made this far without all the great people around me.

First of all, I would like to thank the Danish Council for Strategic Research (now part of Innovation Fund Denmark) who provided the funding for this PhD project.

My greatest thanks goes to my supervisors Alex Landex and Professor Otto Anker Nielsen for their guidance and support throughout the project. Alex is in great part responsible for luring me into the world of railway research, and this PhD thesis would not have existing without him.

In the spring of 2015 I went on an external research stay at Erasmus University in Rotterdam to visit Professor Leo Kroon and Assistant Professor Marie Schmidt. They provided very valuable feedback and suggestions on the work on my capacity models which definitely resulted in higher quality. For that I am very grateful. An extra thanks also goes to Marie Schmidt for taking me out in my spare to teach me her great passion – climbing.

Thanks also to Troels Richter and Bernd Schittenhelm from Rail Net Denmark (Banedanmark) for providing delay data, and timetable and infrastructure data, respectively, for the work on robustness indicators. For my work on the capacity models, I also thank Thomas Rasch for his ideas on how to mix trains and giving valuable feedback in the beginning of the PhD project.

Furthermore, I want to thank Ulla Steen Salado-Jimena for proofreading the thesis and papers during the course of the PhD project.

A lot of thanks also goes to my many great colleagues and persons whom I met at conferences and seminars. Both for scientific discussions, but also for a lot of fun as well as giving me a pleasant work environment.

Lastly, but obviously not least, I send my sincere gratitude to my family, especially my parents and sister, my friends, and of course my dear girlfriend Marie Askholm for the endless support in times of frustrations as well as taking my mind of the PhD project for other activities.

Thank you,
Lars Wittrup Jensen

Summary

In a world continuous striving for higher mobility and the use of more sustainable modes of transport, there is a constant pressure on utilising railway capacity better and, at the same time, obtaining a high robustness against delays. During the planning of railway operations and infrastructure this can be assisted by improving decision support systems to enable planners to use their time more efficiently. In the context of strategic (long-term) planning, efficient decision-support tools translate into being able to evaluate infrastructure and timetable scenarios fast with little data input.

This has motivated the research conducted and described in this thesis, where the objective has been to develop and improve existing methods to achieve timetable and infrastructure plans with robust capacity utilisation aimed at the strategic and early tactical planning phases. Specifically, this thesis contributes with the following main contributions:

- Development of timetable complexities as a new robustness indicator
- Evaluation of selected robustness indicators based on a comparison with microscopic simulations
- Development of a framework, and a model implementing this framework, to assess capacity consumed by a given set of trains (plan of operation) in a network without a predefined timetable
- Development of an optimisation framework to estimate the number of trains a railway network can handle
- A quantitative analysis of relationships between maximum and practical capacity, as well as heterogeneity, capacity, and robustness in networks

The thesis is composed of two parts containing these main contributions on robustness indicators and capacity models, respectively, with the research chapters based on Jensen and Landex (2013a); Landex and Jensen (2013); Jensen et al. (2014, 2015); Jensen (2015)

and Jensen et al. (2016). Additionally, the thesis makes an introduction into planning and modelling of railway systems, as well as reliability of railway operation, including definitions in the chapters 2 and 3.

In chapter 3, a review is conducted by this thesis on definitions of robustness, stability, and resilience of railway systems. Bases on this review the following definitions are synthesised by the thesis. A stable timetable is a timetable that can settle from initial delay within a given time period without dispatching measures. Robustness is the ability of a timetable to withstand smaller disturbances and variance in process times with little or no use of dispatching measures. A timetable that is robust and efficient can provide a service that minimises the realised travel times of passengers (and freight) (an extension of the robustness definition above). The resilient timetable can quickly make a transition into a new stable state when larger disruptions occur using dispatching measures.

Part I of this thesis is concerned with robustness indicators, which are indicators that have the ability to capture size and distribution of initial delays, buffer times, and time supplements either directly or indirectly. Robustness indicators are computationally more simple and less data intensive than a simulation and are therefore suitable for optimisation models and early (strategic) planning phases.

In chapter 4, this thesis identifies the following (types of) indicators for robustness of railway systems: headway heterogeneity indices, UIC 406 compression results, distribution of running time supplements, network locations with a high risk of delay propagation, and timetable and infrastructure complexities. It is concluded by this thesis that none of these indicators are able to capture initial delays, buffer times, and time supplements at the same time. However, separately, the indicators are able to capture some of these parameters. For instance, the UIC 406 method can indirectly be used to indicate the amount of (minimum) buffer times available. While headway heterogeneity measures can be used as an indicator for distribution of buffer times.

In relation to stations where there is a high risk of delay propagation due to train interdependencies, three methods on infrastructure complexities have been reviewed by this thesis in chapter 4. These complexity index methods are used to reflect possible route conflicts at a station with an increasing level of detail. This thesis finds that these methods has some ability to indicate the risk of delay propagation at stations. However, this thesis also finds that they may be improved by taking the exact timetable into account. This thesis therefore develops a timetable complexity index. Based on the infrastructure, timetable, and a probability distribution for delays, the complexity of the timetable and station is calculated. This complexity indicator can indicate robustness through the ability

to capture distribution and size of buffer times and train delays, which is confirmed in a case study conducted on the Danish station Skanderborg.

The indicators reviewed or developed in chapter 4 can measure the robustness of timetables either directly or indirectly based on the distribution and size of initial delays, time supplements, and buffer times from a theoretical point of view. To reveal the practical usefulness of the indicators, microscopic simulations have been conducted in chapter 5 on the North West Railway line in Denmark. Subsequently, this thesis compared simulation results with robustness indicator results for the following indicators: headway heterogeneity, infrastructure and timetable complexity indices, UIC 406, and train path risk profiles. From the comparison conducted, this thesis concludes that the most detailed infrastructure complexity index, the timetable complexity index, the UIC 406 results, and the train path risk profiles perform best in terms of capturing robustness changes.

Part II describes the development of methods to assess capacity in networks. Initially, chapter 6 illustrates that railway capacity is not easily defined as it depends on several infrastructure, rolling stock, and operation parameters. A review of methods to assess capacity reveals that no existing method, spanning over queuing models, optimisation approaches, simulation and sampling models, and many others, can assess capacity in network without a predefined timetable while considering robustness against delays at the same time.

In chapter 7, this thesis therefore develops a framework for the assessment of infrastructure occupation and capacity consumption in railway networks in the strategic planning phase. The framework outlines the overall procedure which may be implemented in any way found suitable. The framework extends current methods, especially the widely used UIC 406 method, to consider both networks and robustness without a predefined timetable. Only the plan of operation (service intentions) in the form of the number and characteristics of trains as well as the infrastructure layout and headway times are needed.

Infrastructure occupation values are derived in chapter 7 using an adjusted longest path search based on the minimum headway times required between trains. To obtain the capacity consumption, this thesis develops and uses a stochastic simulation to estimate buffer times (to account for robustness against delays) based on the delay propagation caused by a set of sampled input delays.

The developed framework, and the model implementation of it, is used on a case network in Southwestern Denmark with two at-grade junctions and four infrastructure scenarios. In all scenarios, the model developed by this thesis can quantity improvements in network infrastructure occupation and capacity consumption. In addition to assessing

infrastructure alterations, the impact on capacity by rolling stock changes and alternative line plans (service intentions) may also be assessed using the developed approach.

Furthermore, results from the case show that the results of capacity assessments are very much dependent on the size of the network considered. This is a consequence of adding more constraints to the calculation as the network grows. Therefore, this thesis suggests to consider the whole network and to evaluate capacity consumption rather than infrastructure occupation to include the aspect of robustness.

In chapter 8, this thesis develops an optimisation framework for the estimation of capacity in terms of number of trains a railway network can handle. The developed approach makes it possible to determine the number of times a certain mix of train can be scheduled in a network plus additional trains that can be added according to predefined train type weights.

The developed optimisation framework takes the heterogeneity of operation into account, contrary to many existing approaches, and is thus a powerful tool for early (strategic) planning phases as it is able to give a span of capacity that a future timetable may utilise. Furthermore, the developed optimisation framework provides the means to ensure that the estimated capacity can yield an operation robust against delays if fully utilised (and timetabled properly).

The developed optimisation framework can estimate the capacity for the network in Southwestern Denmark in 3-5 minutes. Thus for strategic planning the model runs fast and can therefore also be used for larger test instances.

In chapter 9, using the developed model of chapter 7, this thesis investigates the relationship between practical capacity and maximum capacity in networks, and the relationship between speed heterogeneity, infrastructure occupation, and robustness in networks.

In the chapter, this thesis finds that the additional time rates are lower in networks than the recommendations given for line sections by UIC (2004b, 2013), i.e. for one network tested 18.3% compared to 33%. Additional time rates are rates to be added to infrastructure occupation values to derive if the practical capacity has been exceeded. The lower rates are caused by additional buffer times inherently available in compressed network timetables compared to compressed timetables on line sections. This is also observed in different scenarios that show a correlation between train dependencies in the network yielding additional buffer times and thus lower additional time rates.

In chapter 9, the relationship between heterogeneity and infrastructure occupation in networks is also investigated. First a heterogeneity indicator for running time heterogeneity in networks is developed. Subsequently, it is found that for networks the relationship between heterogeneity and infrastructure occupation is best represented by a recipro-

cal function. For simpler network it is found that a linear function also represents this relationship well.

Using the developed heterogeneity index, the relationship between heterogeneity, infrastructure occupation, and robustness (consecutive delays) is investigated by this thesis. It is found that train sequences with low heterogeneity have a risk of higher maximum delay propagation (consecutive delays) than train sequences with higher heterogeneity, as the latter sequences have (more) additional buffer times implicitly added due to their (increased) heterogeneity. Chapter 9 also illustrates how network dependencies may lead to higher consecutive delays (a network effect), for a large proportion of possible train sequences.

Resumé (Danish)

I en verden hvor der konstant stræbes efter bedre mobilitet og brugen af mere bæredygtige transportformer, er der hele tiden fokus på at udnytte jernbanekapacitet bedre og opnå en højere robusthed mod forsinkelser. Under planlægningen af jernbanedrift og infrastruktur kan dette opnås ved at forbedre beslutningsstøtteværktøjer, der muliggør, at planlæggere kan bruge deres tid mere effektivt. I forbindelse med strategisk (langsigtet) planlægning kan effektive beslutningsstøtteværktøjer overføres som værende værktøjer der kan vurdere infrastruktur og køreplansscenarier hurtigt med minimalt input.

Dette har motiveret forskningen beskrevet i nærværende afhandling. Det har heri været målet at udvikle og forbedre eksisterende metoder til at opnå køre- og infrastrukturplaner med en robust kapacitetsudnyttelse rettet mod den strategiske og tidlige taktiske planlægningfase. Denne afhandling bidrager konkret med følgende hovedbidrag:

- Udvikling af køreplanskompleksiteter som en ny robusthedsindikator
- Evaluering af udvalgte robusthedsindikatorer baseret på en sammenligning med mikroskopiske simuleringer
- Udvikling af et framework, og en model der implementerer dette framework, til at vurdere kapaciteten der forbruges af en given mængde tog (driftsoplæg) i et netværk uden en foruddefineret køreplan
- Udvikling af en optimeringsmodel til estimering af det antal tog et jernbanenetværk kan håndtere
- En kvantitativ analyse af forholdet mellem maksimum og praktisk kapacitet, samt heterogenitet, kapacitet og robusthed i netværk

Afhandlingen består af to dele omhandlende henholdsvis robusthedsindikatorer og kapacitetsmodeller primært baseret på Jensen and Landex (2013a); Landex and Jensen (2013); Jensen et al. (2014, 2015); Jensen (2015) og Jensen et al. (2016). I afhandlingen

introduceres derudover planlægning og modellering af jernbanesystemer i kapitel 2, samt pålideligheden af jernbanedrift, herunder også definitioner, i kapitel 3.

Kapitel 3 indeholder et studie af definitioner af robusthed, stabilitet og modstandsdygtighed (resilence) af jernbanesystemer. I henhold til dette studie kan de følgende definitioner opstilles. En stabil køreplan er en køreplan, der stabiliserer sig ved indledende forsinkelser inden for en veldefineret periode uden aktiv indgriben fra fjernstyringen. Robusthed er en køreplans formåen til at kunne modstå mindre forstyrrelser og varians i procestider med ingen eller minimal indgriben fra fjernstyringen. En køreplan der er robust og effektiv, kan yde en service, der minimerer de realiserede rejsetider for passagerer (og gods) (en udvidelse af definitionen af robusthed herover). Den modstandsdygtige (resilience) køreplan kan hurtigt ændres for at opnå en ny stabil tilstand, når større driftsafbrydelser opstår ved brug af aktiv indgriben fra fjernstyringen.

Del I i denne afhandling omhandler robusthedsindikatorer, som er indikatorer der har evnen til at opfange størrelse og fordeling af indledende forsinkelser, buffertider og køreplanstillæg, enten direkte eller indirekte. Robusthedsindikatorer er beregningsmæssigt mere enkle og mindre dataintensive end simulering og er derfor velegnet til optimeringsmodeller og tidlige (strategiske) planlægningsfaser.

I kapitel 4 gennemgår denne afhandling følgende (typer af) indikatorer for robuste jernbanesystemer: togfølgeheterogenitet, UIC 406, fordeling af køretidstillæg, steder i et netværk med en høj risiko for spredning af forsinkelse samt køreplan- og infrastrukturkompleksiteter. Det konkluderes i afhandlingen, at ingen af disse indikatorer er i stand til at opfange indledende forsinkelser, buffertider og køretidstillæg på samme tid. Hver indikator er dog i stand til at opfange nogle af robusthedsparameterene. For eksempel kan UIC 406-metoden indirekte anvendes til at angive mængden af (minimum) buffertider til rådighed. Ydermere kan togfølgeheterogenitet anvendes som en indikator for fordeling af buffertider.

I forbindelse med stationer, hvor der er en stor risiko for spredning af forsinkelse grundet indbyrdes tog- og ruteafhængigheder, er tre metoder, omhandlende infrastrukturkompleksiteter gennemgået, i kapitel 4. Disse kompleksitetindeksmetoder anvendes til at afspejle mulige rutekonflikter på en station med en stigende detaljeringsgrad. Af afhandlingen konkluderes det, at disse kompleksitetsindekser kan bruges til at estimere risikoen for spredning af forsinkelse på stationer. Det findes dog også, at de kan forbedres ved at tage køreplanen i betragtning. I denne afhandling udvikles derfor et køreplanskompleksitetsindeks. Baseret på infrastrukturen, køreplanen og en sandsynlighedsfordeling for forsinkelser kan kompleksiteten for en given køreplan beregnes. Denne

kompleksitetsindikator kan bruges til at indikere robusthed baseret på fordelingen og størrelsen af buffertider og forsinkelser.

Indikatorerne opstilllet i kapitel 4 kan måle robustheden af køreplaner, enten direkte eller indirekte, baseret på fordeling og størrelse af indledende forsinkelser, køretidstillæg, og buffertid fra et teoretisk synspunkt. For at afdække den praktiske anvendelighed af disse indikatorer, er der i kapitel 5 foretaget mikroskopiske simuleringer med Nordvestbanen som case. Simulationsresultaterne er efterfølgende sammenlignet med resultater for følgende robusthedsindikatorer: togfølgeheterogenitet, infrastruktur- og køreplanskompleksitetindekser, UIC 406 og risikoprofiler. Fra sammenligningen konkluderes det, at køreplanskompleksitetsindekset, UIC 406-resultater, risikoprofiler og det mest detaljerede infrastrukturkompleksitetindeks præsterer bedst i forhold til at opfange robusthedsændringer.

Del II beskriver udviklingen af modeller til vurdering af kapacitet i jernbanenetværk. Indledningsvis illustrerer kapitel 6, at jernbanekapaciteten ikke kan defineres let, da kapacitet afhænger af flere parametere fordelt på infrastruktur, rullende materiel, og driftkarakteristika. En gennemgang af metoder til at vurdere kapaciteten afslører, at ingen eksisterende metoder, herunder bl.a. kø-, optimerings- og simuleringsmodeller, kan vurdere kapaciteten i netværket uden en foruddefineret køreplan mens robusthed mod forsinkelser samtidig tages i betragtning.

I kapitel 7 udvikles derfor et framework til vurdering af kapacitetsforbrug i jernbanenetværk. Frameworket skitserer den overordnede procedure, og kan således benyttes med hvilken som helst metode der findes passende. Frameworket bygger videre på eksisterende metoder, især den velkendte UIC 406-metode, så både netværk og robusthed tages i betragtning uden en foruddefineret køreplan. Som input er det kun nødvendigt med et driftsoplæg med serviceintentioner i form af antal og karakteristika af tog samt infrastrukturlayout og togfølgetider.

Værdier for kapacitetsforbrug er udledt i kapitel 7 ved hjælp af en justeret longest path-algoritme baseret på minimumstogfølgetider. For at opnå kapacitetsforbruget (inklusiv kritiske buffertider - capacity consumption) udvikles og anvendes en stokastisk simulering til at estimere buffertider (for at tage højde for nødvendig robusthed mod forsinkelser) baseret på følgeforsinkelser forårsaget af et sæt af indledende forsinkelser.

Det udviklede framework, og implementering af dette, bruges på et netværk i Sydvest Danmark med to udfletninger i niveau, hvor fire infrastrukturscenarier undersøges. I alle scenarier kan modellen, udviklet i denne afhandling, måle forbedringer i netværkskapacitet og øget robusthed mod forsinkelsesspredning. I tillæg til at vurdere infrastrukturændringer, kan modellen også vurdere følgerne af ændringer i rullende materiel samt alternative linjeplaner for infrastrukturkapaciteten.

Desuden viser resultater fra casestudiet, at kapacitetsestimater er meget afhængige af størrelsen af det betragtede netværk. Dette er en konsekvens af yderligere begrænsninger tilføjet til beregningen i takt med at netværket vokser. Derfor foreslås det, i denne afhandling, at hele netværket betragtes og i denne sammenhæng at kapacitetsforbruget inklusiv kritiske buffertider (capacity consumption) vurderes snarere end uden kritiske buffertider (infrastructure occupation), således at der tages højde for robusthed.

I kapitel 8 udvikles en optimeringsmodel til at vurdere kapaciteten i form af antallet af tog et givent jernbanenetværk kan håndtere. Den udviklede model gør det muligt at bestemme antallet af gange et bestemt miks af tog kan håndteres af et netværk samt et ekstra antal tog, der maksimalt kan tilføjes i henhold til foruddefinerede togtypevægtninger.

Optimeringsmodellen tager højde for heterogeniteten af driften, i modsætning til mange eksisterende tilgange, og er således et godt værktøj i tidlige (strategiske) planlægningfaser, hvor det kan give et spænd af kapacitet, som en fremtidig køreplan kan udnytte. Desuden giver optimeringsmodellen mulighed for at sikre, at den estimerede kapacitet kan give en drift, der er robust imod spredning af forsinkelser, selv hvis den udnyttes fuldt ud (og køreplanlægges omhyggeligt).

Optimeringsmodellen kan estimere kapaciteten for netværket i Sydvest Danmark på 3-5 minutter. Således er modellen beregningsmæssig hurtig i et strategisk planlægningsperspektiv, og der er derfor gode perspektiver i også bruge modellen til større netværk.

I kapitel 9 undersøges forholdet mellem praktisk og maksimal kapacitet i netværk, samt forholdet mellem hastighedsheterogenitet, kapacitetsforbrug og robusthed i netværk ved hjælp af den udviklede model fra kapitel 7.

I kapitel 9 konkluderes det, at den nødvendige tillægssats er lavere for netværk i forhold til anbefalingerne for strækningsafsnit givet af UIC (2004b, 2013). For det testede netværk findes for eksempel en værdi på 18,3 %, hvilket skal ses i forhold til UIC 406-anbefalingen for strækningsafsnit på 33 %. Tillægssatser er rater der skal ganges på kapacitetsforbruget for at udlede hvor vidt den praktiske kapacitet er overskredet. De lavere tillægssatser, der er nødvendige, er forårsaget af ekstra buffertider, der er tilgængelige i komprimerede netværkskøreplaner i forhold til komprimerede køreplaner på strækningsafsnit. Dette er observeret i forskellige scenarier, der viser en sammenhæng mellem togafhængigheder i netværket og færre nødvendige tillægssatser (grundet større buffertider).

I kapitel 9 er forholdet mellem heterogenitet og kapacitetsforbrug i netværk ydermere undersøgt. Først udvikles en heterogenitetsindikator for køretid i netværk. Efterfølgende konkluderes det, at forholdet mellem heterogenitet og kapacitetsforbrug for netværk bedst

repræsenteres ved en reciprok funktion. For enklere netværk konkluderes det tillige, at forholdet mellem de to parametre afspejles godt af en lineær funktion.

Med udgangspunkt i det udviklede heterogenitetindeks er forholdet mellem heterogenitet, kapacitetsforbrug og robusthed (forsinkelsesspredning) undersøgt i denne afhandling. Det konkluderes, at togfækkefølger med lav heterogenitet har en risiko for højere maksimal spredning af forsinkelse (følgeforsinkelser) end togfækkefølger med højere heterogenitet, hvor sidstnævnte togrækkefølger har (større) buffertider implicit indbygget på grund af disses (forøget) heterogenitet. Derudover illustreres det også i kapitel 9, hvordan netværksafhængigheder kan føre til højere følgeforsinkelser for en stor del af de mulige togrækkefølger.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

In recent years there has been an increasing focus on achieving a more environmentally friendly transport sector. This has resulted in the EU white paper "Roadmap to a Single European Transport Area" from 2011 (European Commission, 2011) and "A green transport policy" by the Danish Parliament in 2009 (Danish Ministry of Transport, 2009). Both are calling for a shift to more sustainable transport modes. A natural step to obtain this shift in transport mode is to move more passengers and freight to railways. This in turn calls for more capacity and/or higher capacity utilisation. Within railway planning it is well-known that increased capacity consumption leads to increased risk of delays and thus a less robust system. As passengers value the reliability of train services highly, to achieve a shift to rail and at the same time increase or maintain reliability and robustness against delays is thus a challenge.

1.1 Motivation

Planning railway operations and infrastructure is a complex task subject to many operational and infrastructure constraints as well as external factors and requirements. Conceiving infrastructure and operation plans to provide cost-efficient, optimally utilised, and highly available train services is therefore difficult. However, one way to facilitate such plans is the use of decision-support tools. Such tools enable planners to plan infrastructure and operations more time efficiently as considerably more solutions can be evaluated compared to manual planning.

A common decision-support tool to evaluate robustness of infrastructure and operational plans against delays is the use of microscopic simulation. It is accurate and offers a high level of detail, however at the cost of considerable amounts of input data, long set-up time, and long computation times.

In long-term planning phases, detailed data may not be available and the long set-up times of microscopic models are not desirable as many alternatives may have to be evalu-

ated. As an alternative, indicators that, directly or indirectly, indicate the robustness of a railway system based on system characteristics can be used. However, the actual ability of these indicators to express changes in robustness may differ significantly. Furthermore, the semantics of robustness indicators are mainly unknown and indicators are therefore best for comparison of alternatives. Thus there is a need for evaluating current robustness indicators, and possibly to improve these, and develop new indicators.

In the context of strategic (long-term) capacity planning, efficient decision-support tools translate into being able to evaluate capacity scenarios fast with little data input. However, the capacity of a railway system is not easily determined. This is the case as capacity depends not only on the infrastructure, but also on operational constraints, the order of trains (if operating speeds are different), and the rolling stock used (UIC, 1996).

As a consequence, the capacity of a railway system cannot be assessed with traditional methods without a timetable. This is not desirable, as the process of producing a timetable can be very time-consuming. Even if a timetable has been produced, it is uncertain if it will be put into operation due to the long planning horizon. Furthermore, to reduce the work-load in this process only smaller parts of the network are usually considered, thus neglecting constraints induced by other parts of the network (Jensen et al., 2016). Within strategic capacity planning, there is therefore a need to advance current methods to provide high quality estimates of capacity fast and efficiently.

1.2 Objectives

The general objective of this thesis is to develop methods to achieve timetable and infrastructure plans with robust capacity utilisation aimed at the strategic and early tactical planning phases (5 years and earlier before realisation). Specifically, this general objective is divided into two sub-objectives which are aimed at the issues described in the previous section:

1. Review, evaluate, and develop indicators that can indicate the robustness of railway operation and infrastructure plans

2. Develop a methodology for assessment of (robust) capacity in railway networks independent of a predefined timetable but based on service intentions (plan of operation)

1.3 Main contributions

Based on these objectives, the research conducted, and described in this thesis, has produced the following main contributions:

- Development of timetable complexities as a new robustness indicator
- Evaluation of selected robustness indicators based on a comparison with microscopic simulations
- Development of a framework, and a model implementing this framework, to assess capacity consumed by a given set of trains (plan of operation) in a network without a predefined timetable
- Development of an optimisation framework to estimate the number of trains a railway network can handle
- A quantitative analysis of relationships between maximum and practical capacity, as well as heterogeneity, capacity, and robustness in networks

The following two subsections summarise the main contributions listed above in greater detail. In addition to these main contributions, literature reviews have been conducted in this thesis. These reviews provide an overview of definitions of robustness, stability, resilience, and capacity as well as models for robustness analysis and capacity assessment.

1.3.1 Robustness indicators

Based on the first objective of this thesis, the initial focus is on robustness indicators. First, indicators that are potentially able to indicate the robustness of a given railway system are surveyed and identified. Based on this review a timetable complexity indicator is developed that extends existing infrastructure complexities to assess robustness of a timetable at stations.

Selected indicators are subsequently evaluated and compared with a microscopic simulation for their ability to indicate robustness. The indicators can be used in early screening (strategic) phases or as a part of analytical methods and optimisation models for improved computational speed and reduced input compared to microscopic simulations.

1.3.2 Models for assessment of railway capacity

In relation to the second objective, a main contribution of this thesis is the development of a framework to evaluate capacity consumption in railway networks in the strategic planning phase. The framework is characterised by the following:

- No predefined timetable is needed (only the service intentions in a plan of operation)
- Networks can be handled, not just lines or line sections
- Estimation of buffer times for absorbing delays and thus achieving a robust capacity utilisation

Within this framework, it is proposed that multiple permutations of train orderings (sequences) of a given set of trains are evaluated to produce a span (output distribution) of infrastructure occupation. This deals with the problem of infrastructure occupation and capacity consumption depending on heterogeneity, and consequently a timetable is not needed for evaluation. Only the plan of operation in the form of number of trains and their characteristics is needed (routes, headways, stop patterns, and running times). This reduces the work-load in the strategic planning phase, where the model can be used to evaluate capacity ramifications of strategic decisions. Furthermore, the thesis proposes to evaluate the whole network and not just line divisions, as these divisions (due to network effects) can lead to ambiguous results as in UIC (2004b, 2013) .

To implement this framework, a model is developed and demonstrated. The model developed is based on a mesoscopic infrastructure model. The model generates a number of compressed timetables for a given network and set of trains using an adjusted longest path search from where the infrastructure occupation can be derived. Using a discrete event simulation, the compressed timetables are simulated to estimate critical buffer times and produce the capacity consumption.

The developed model is used on a case network to show how different infrastructure scenarios can be evaluated based on the same plan of operation. Furthermore, the thesis compares and discuss network infrastructure occupation and capacity consumption values with the line division recommendations by UIC (2004b, 2013), and thus the relation between maximum and practical capacity for networks. Additionally, relationships between heterogeneity, infrastructure occupation, and robustness in networks are investigated using the model developed.

Based on the above-mentioned framework and model, the thesis also contributes with the development of an optimisation framework to determine the number of trains that can enter a given network under robust capacity utilisation.

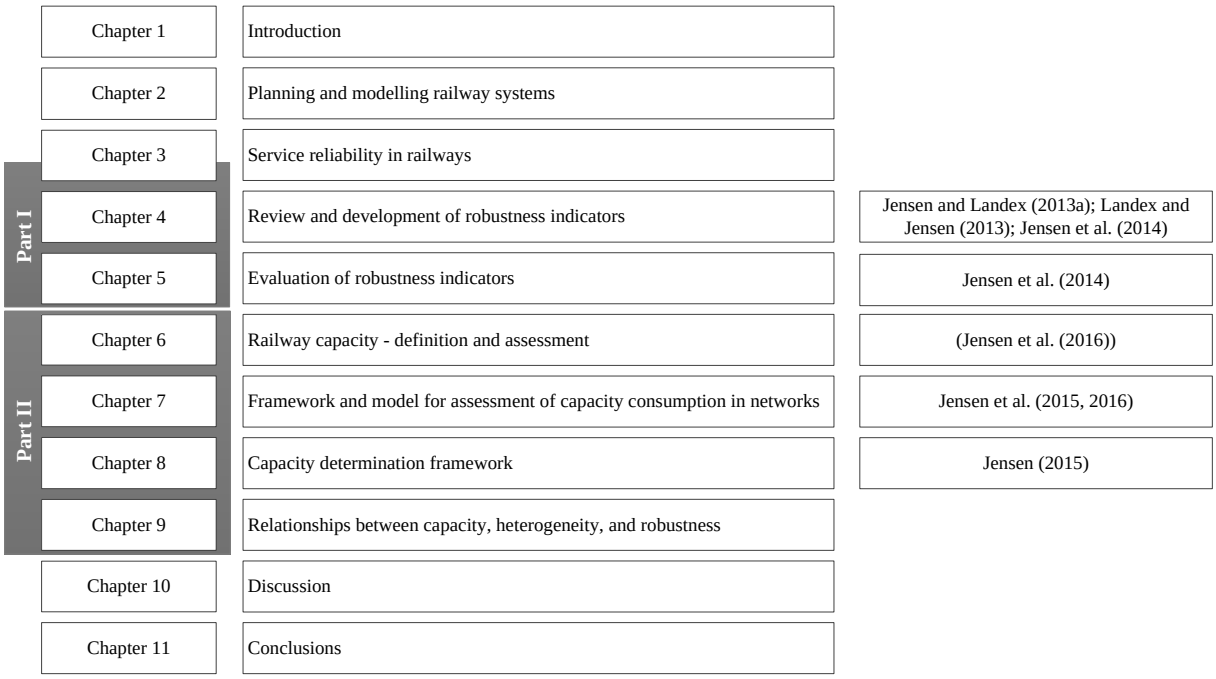


Figure 1.1: Thesis structure with references to papers that chapters are based on.

1.4 Thesis outline

This thesis consists of 12 chapters with the research chapters divided into two parts on robustness indicators and capacity modelling, respectively. The research chapters are based on the articles: Jensen and Landex (2013a); Landex and Jensen (2013); Jensen et al. (2014, 2015); Jensen (2015); Jensen et al. (2016) as well as work not published yet. The structure of the thesis, as well as the references to the mentioned papers, is depicted in figure 1.1. Both parts can be read individually, therefore part I does not necessarily have to be read before part II.

Besides this chapter, chapters 2 and 3 serve as an introduction and background for both part I and II. Specifically, chapter 2 is an introduction to planning and modelling of railway systems. This provides a short description of essential models for timetable and infrastructure planning, the terminology used as well as the context of the contributions made by this thesis. Chapter 3 gives an introduction to concepts, terminology, and assessment methods related to reliability of railway services. In particular, the definitions of stability, resilience, and robustness of railway timetables are reviewed. Chapter 3 serves as an essential background chapter for part I and should, therefore, be read prior to part I.

Part I of this thesis consists of chapters 4 and 5 on robustness indicators. First chapter 4 gives a comprehensive review of robustness indicators. Furthermore, a new

robustness indicator in the form of timetable complexities is developed by this thesis are described in chapter 4. Based on chapter 4, selected robustness indicators are used and demonstrated on a regional railway line on Zealand and compared with results of a microscopic simulation in chapter 5. Based on this comparison, it is evaluated how the selected indicators are able to capture robustness.

Part II contains four chapters related to capacity modelling of railways. First, chapter 6 reviews existing methods to assess capacity as well as give an introduction to the concept, definition, and challenges in assessing railway capacity. Chapter 7 describes the framework and model developed by this thesis for strategic assessment of capacity consumption in railway networks based on a plan of operation. Subsequently, in chapter 8, an optimisation framework is developed and tested. This framework extends the work in chapter 7 to estimate the number of trains that a given network can handle. Finally, in chapter 9 the model described in chapter 7 is used to explore the relation between practical and maximum capacity in networks. Furthermore, relationships between infrastructure occupation, speed heterogeneity, and robustness are investigated.

Chapters 10 and 11 complete this thesis. First, in chapter 10, the findings and developed methodologies are discussed and directions for further work are suggested. Lastly, conclusions of the research described in the thesis, is given in chapter 11.

2 Planning and modelling railway systems

In this chapter an introduction is given to modelling and planning of railway systems as well as the terminology used in the context of operations planning. The chapter serves as an overview for readers not familiar with modelling and planning of railway systems. For readers familiar with this topic, the chapter provides the terminology used throughout the thesis.

2.1 Planning phases

The railway operational process can be divided into several sub-processes which serve to solve the complex task of planning the railway operations by decomposition. In the operations planning process of public transit systems four activities can be listed according to Ceder (2007): network route design, timetable development, vehicle (rolling stock) scheduling, and crew scheduling. Further processes related to long and short-term planning can be added to this as depicted in figure 2.1. It is common to divide the sub-processes into a strategic, a tactical level, and an operational level based on their planning horizon (Kaas, 1998b; Caimi, 2009; Lusby et al., 2011).

The strategic planning phase usually takes place 5 years or more before plans are put into operation (Kaas, 1998b; Caimi, 2009). Strategic planning includes the planning of infrastructure, rolling stock procurement, and line planes. In line planning, the future train services are planned in terms of stops served by a service and the frequency of services based on passenger demand. The infrastructure, rolling stock and the line plan form the basic components for the timetable and the crew and vehicle schedules. Larger infrastructure changes and rolling stock purchases can take considerably more than 5 years. Up to 20 years is not unusual in this regard (Caimi, 2009).

The tactical planning phase takes place a couple of years to a couple of months before the timetable and vehicle and crew schedules are put in to operation (Kaas, 1998b). At the Danish infrastructure manager Rail Net Denmark, the tactical planning of the timetable

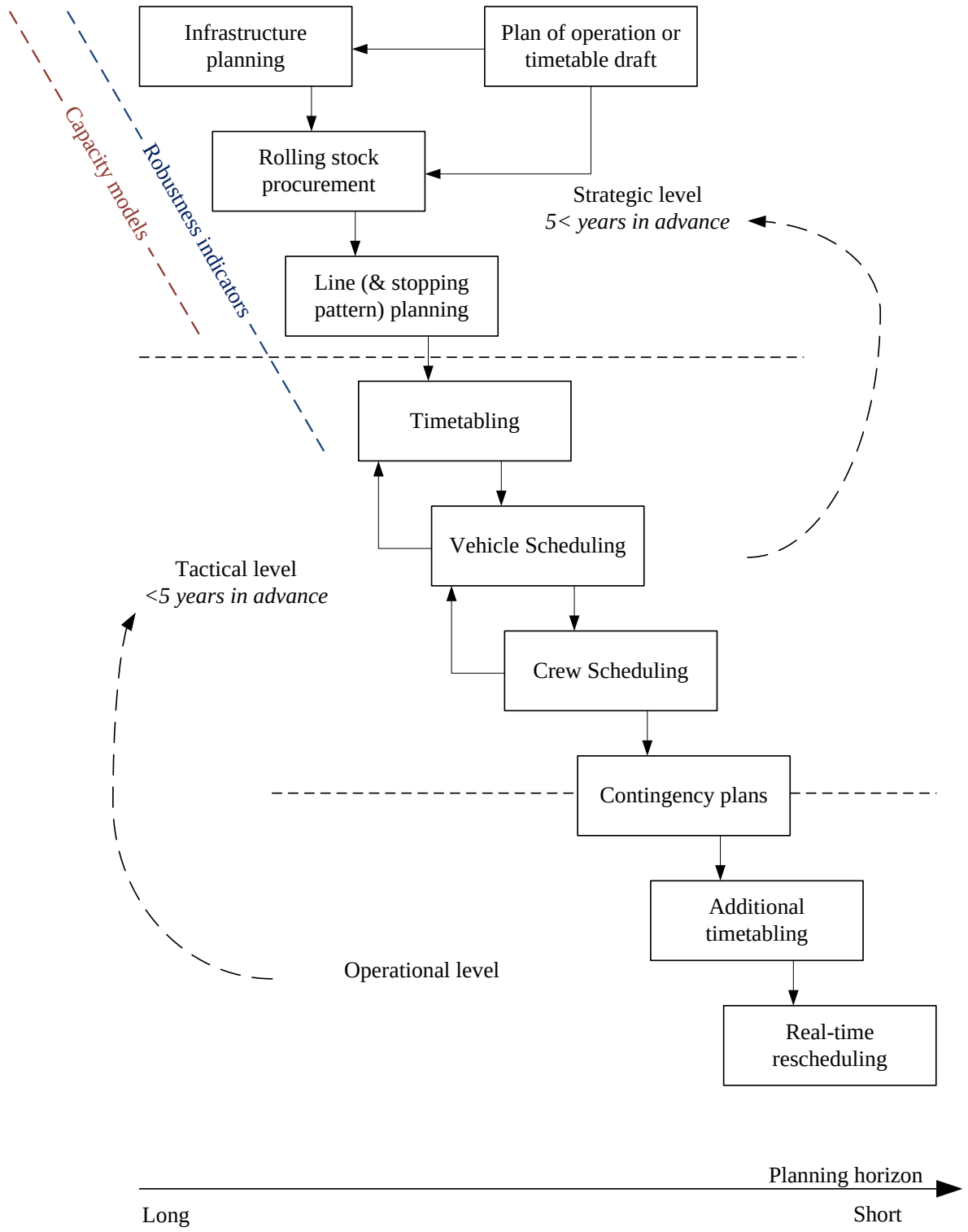


Figure 2.1: The planning process in railways. Contributions of this thesis marked with red and blue.

is started 4 years before it is put into operation (Schittenhelm, 2013). During the tactical planning phase a detailed timetable is derived based on the available infrastructure, the rolling stock, as well as passenger and freight demand (Schittenhelm, 2013). The timetable may be derived in two different ways, either completely by a company managing the infrastructure and the rolling stock or as an negotiation between an infrastructure manager and one or more train operating companies. These companies may either be private or publicly owned. A division into infrastructure manager and train operating company has been dictated for the EU countries with the "rail infrastructure package" in 2001 consisting of three EU directives (European Commission, 2012). This legislation has been passed to obtain non-discriminatory access to the railway infrastructure for all train operators to increase integration of the internal market of the EU. Based on the timetable derived, either with or without negotiation, detailed rolling stock and crew schedules can be planned by each operator.

On the boundary between the tactical and operational phase, contingency operation may be planned based on the timetable and crew and rolling stock schedules, as well as a number of disruption scenarios for quick re-scheduling in the operational phase.

In the operational phase, the operation of railway system is carried out according to the plans in the tactical phase. Remaining capacity may be allocated on an ad-hoc basis (Schittenhelm, 2013), e.g. extra freight trains. Large delays in traffic may be dealt with using contingency plans and real-time rescheduling of trains, crew and rolling stock.

Each of these planning processes can be solved either manually or by the aid of computer models. Due to the complexity of the planning process, integrated planning of all phases is not possible (Ceder, 2007). Even using computers, exact solution procedures for an optimised integrated plan are not to be expected (Schöbel, 2011). The different sub-processes are therefore carried out sequentially instead of simultaneously, where the output of one sub-process is the input to the next (Lusby et al., 2011; Ceder, 2007).

To improve the output of the different sub-processes some information of later processes are necessary. Thus feedback are usually given gradually as indicated in figure 2.1. In practice this is done between planners of different phases/processes, but can also be the result of passenger surveys and passengers counting (van Oort, 2011) as well as through the use of delay data from the operation.

Furthermore, for some sub-processes input is necessary from sub-processes that are placed later in the planning process. An example of this is timetable planning. If rolling stock and crew cannot be feasibly assigned to the planned timetable, it is not possible to operate the timetable in practice. Thus simple schedules of rolling stock and crew assignment have to be drafted during the timetable planning process. Detailed and optimised

rolling stock schedules can then subsequently be derived based on the detailed timetable. In early strategic planning phases it is often also necessary to work out draft timetables or plans of operation based on future demand when planning larger infrastructure expansions and procurement of rolling stock.

Beyond simple feedback loops and more formal iterative procedures, there is an increasing focus to obtain better plans by integrating two or more planning processes by using advanced operations research techniques and powerful computers. E.g. see Schöbel (2011) for an overview of literature associated with the integration of line planning with other planning steps.

The contributions of this thesis and their relation to the planning phases are shown in figure 2.1. Robustness indicators are dealt with in chapters 4 and 5, and relates to infrastructure planning, remotely to line planning, and especially to timetabling. The capacity models proposed by this thesis in chapters 7 and 8 are mainly suitable for infrastructure planning, but can also be used in relation to rolling stock procurement and line planning for specific use cases. See the referenced chapters for more details.

2.2 Modelling railways

With computational power always on the rise there is a huge incentive to use computers to aid planners in obtaining the best plans possible. Abstractions of the railway system have to be done to use computer aided planning systems. In this section, an introduction is therefore given to the components of a railway system and how they can be modelled.

2.2.1 Terminology

The following main components are of interest when planning timetables, crew and rolling stock:

Rolling stock (trains) carry passengers and/or freight. Can either consist of a locomotive with a number of coaches (alternatively denoted cars) or trainsets where the traction equipment is built into the coaches. Different rolling stock have different dynamics (acceleration and deceleration capabilities) and thus running times.

The capability of the **tracks**, ballast and sub-ballast to carry a train with a given axle load under a maximum allowed speed. In curves the allowed speed may be decreased from the generally allowed speed to prevent derailment or the train overturning. Furthermore, the grade at which the tracks are constructed and the resistance in curves affect the train performance and thus the running time of trains.

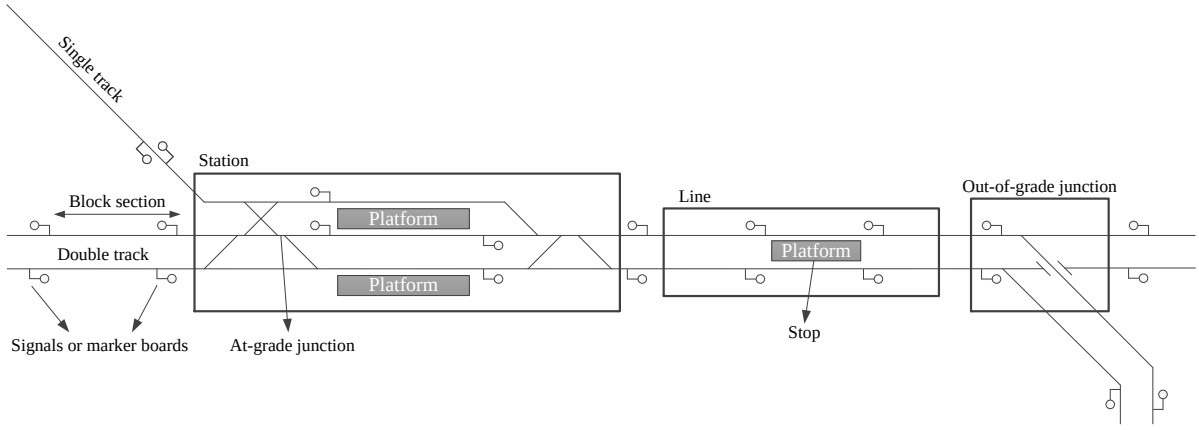


Figure 2.2: Railway network topology.

Switches allow for routing of trains. Also denoted points or turnouts. The diverting branch of a switch usually has to be traversed at a lower speed than the main branch.

The **signalling system** consists of signals, train detection equipment, switch control, and in modern systems ATP (automatic train protection) to allow for the safe and efficient operation of trains. The layout and technology behind a given signalling system are a significant factor for the minimum allowed headway between trains.

Passenger flows and platform layout, the amount of passengers and how they are distributed in time and space govern the dwell time required as well as the required capacity of the trains.

Processes and rules. Departure processes, rules for safe operation, union rules, etc. dictate how the infrastructure, rolling stock and crew have to be used.

As depicted in figure 2.2, the topology of a track network can be divided into the following. **Line sections** defined as sections of the network where no switches are present to allow for (re)routing of trains. **Junctions** where lines meet and switches allow for routing of trains between different lines. In this relation, this thesis distinguish between at-grade and out-of-grade junctions. At-grade junctions have no constructions to allow for avoidance of crossing conflicts between lines. On the contrary out-of-grade junctions have been constructed to avoid crossing conflicts. **Stations** are defined as a place of passengers exchange as well as a place for routing of trains through the use of switches. At larger stations switches are put into place to allow for more platform tracks and (re)routing. At smaller stations, switches are placed to allow for re-routing of trains in contingency

operation or as an overtaking facility. On single track lines switches facilitate passing loops. Stations may also contain junctions. If a station is defined as a terminal station, some or all train services end here. **Stops** are defined as passenger exchange points on the open line. A **railway line** is a section of a network defined between two stations, possibly with intermediate stations. A railway line may be defined as a single, double or multiple track railway line. Railway lines cannot overlap and the union of all railway lines constitutes the network. The term should not be confused with a train service line as related with the line planning process described briefly in section 2.1.

2.2.2 Infrastructure models

Using graph theory, an abstraction of the railway system can be obtained for use in computer models. This abstraction serves as the basis for modelling railway systems in terms of train performance, infrastructure performance, and passenger behaviour. In the following the focus will be on infrastructure modelling with varying levels of detail.

A high level of detail can be obtained by using a microscopic graph structure to represent the infrastructure, also defined as a microscopic infrastructure model. In a microscopic infrastructure model all elements are represented by a vertex or an edge (as depicted in figure 2.3). Each edge defines a track segment with a specific allowed speed, grade, curve resistance, bidirectional operation possible and possibly more. Vertices define switches, stop boards, track circuits/axle counters and signal/marker board placements. Microscopic models have the advantage of being detailed and thus the potential of being very precise. However, they require a significant amount of data and are computationally slow to use due to the many vertices and edges.

To cope with the disadvantages of the microscopic model, a macroscopic infrastructure model can be used (see figure 2.3). In this type of model, a higher level of abstraction is used to reduce the number of edges and vertices. However, this reduction also decreases the precision of the model. Stations are simply represented by vertices and the lines between stations as edges. Data associated with edges are typically running times and minimum headway times, but may also be average speed and capacity in terms of number of trains.

Infrastructure representations using a level of abstraction between macroscopic and microscopic models are defined as mesoscopic infrastructure models (Gille et al., 2010). Such an infrastructure model may use any level of abstraction not as detailed as the microscopic model and not as abstracted as the macroscopic model. In figure 2.3 two

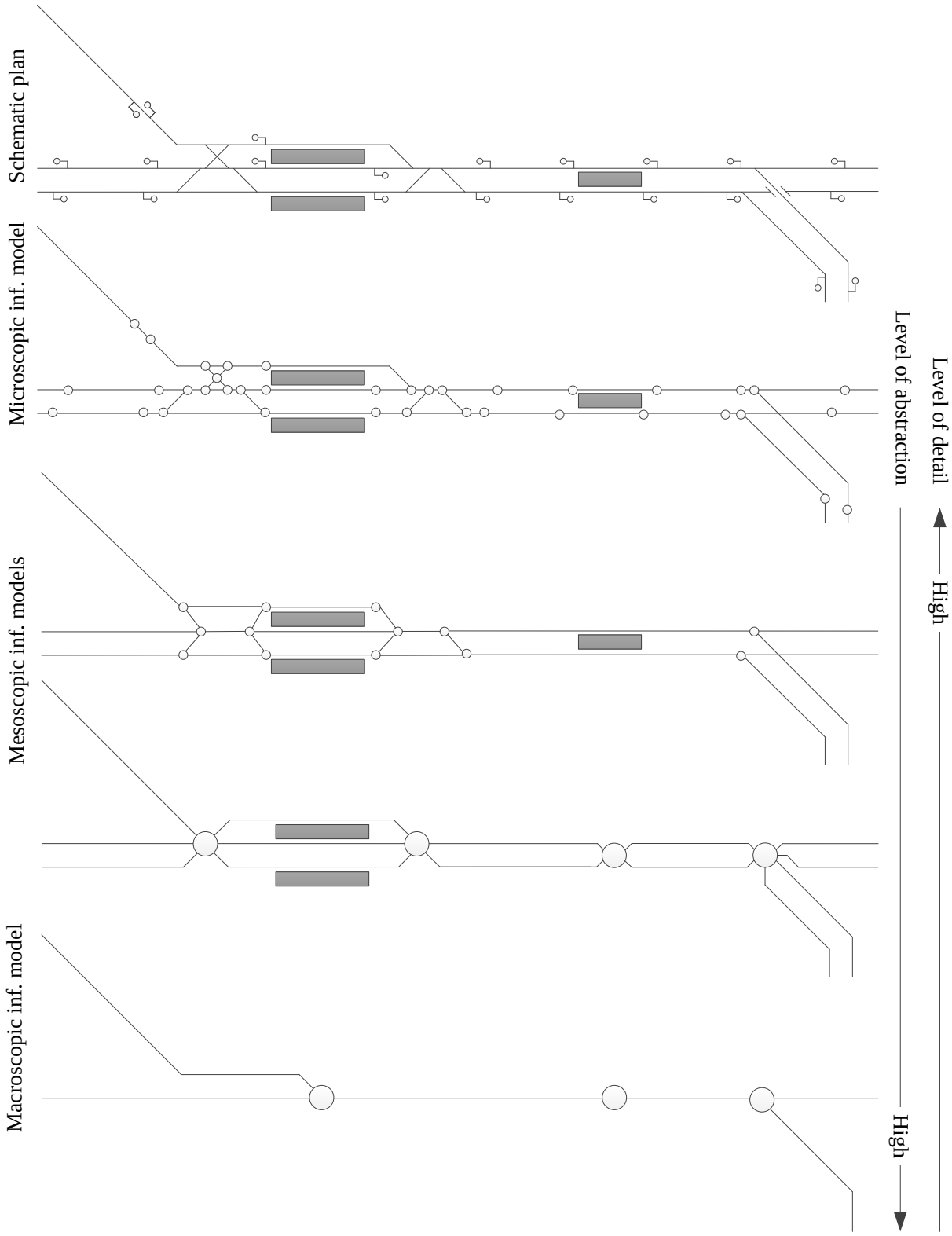


Figure 2.3: Microscopic, mesoscopic, and macroscopic infrastructure models.

different mesoscopic infrastructure models are shown in addition to the macroscopic and microscopic representation.

The use of the microscopic, mesoscopic and macroscopic infrastructure models is often dictated by the planning stages in which they are used due to data availability. Thus in the strategic planning phases only sparse data may be available and a macroscopic model is therefore often the only solution. A microscopic model may be used, but then microscopic data have to be auto-generated based on macroscopic data, only offering little improvement of precision. E.g. Gille et al. (2010) converts a macroscopic model to a mesoscopic model to validate macroscopic results, defined as multiscaling analysis. For the operational phase, the planning horizon is much shorter and high quality detailed data is available making the use of a microscopic model possible.

In addition to the infrastructure model presented above, data on rolling stock and the passengers/freight is necessary to model operation, generate timetables, rolling stock and crew schedules as well as model passenger route choices.

To model train running times, train dynamics of the rolling stock is necessary. For the modelling of dwell times several parameters are needed: number of passengers, number and distribution of doors in the train, length of the train, access points on the platform, width of the platform as well as obstacles on it, and the time since the last train. Given train running times, dwell times as well as operational rules together with reaction time of the signalling system, the minimum allowable headways times can be modelled. This again serves as a basis for timetabling which is based on on a line plan that has been generated based on assumed passenger flows and estimated track capacity. However, as the timetable affects the way passengers choose their route and the number of passengers on the platforms, there is circular dependency between the timetable and the passenger flows. Usually, this is handled by the feedback loops described in section 2.1. E.g. if dwell times are too short or too long they can be adjusted in the next revision of the timetable or, if possible, already in the operational phase.

As stated earlier the disadvantage with macro and mesoscopic infrastructure models is the lack of detail, while the increased precision of the microscopic model comes at the cost of increased computation time. When microscopic data is available and fast computation is needed it is possible to combine a micro and a macro/meso model. Also called the micro-macro model approach.

This can be done in two ways: one-way conversion and two-way conversion. Using a one-way conversion, headway and running times are calculated based on the microscopic model and aggregated to data for the macro/mesoscopic model. This ensures feasibility at the track (micro) level as long as preconditions for the running time and headway

Infrastructure model	Suitable/necessary for
Microscopic	Headway time calculation, running time calculation, realistic simulation of operation, routing, rescheduling
Mesoscopic	Between microscopic and macroscopic
Macroscopic	Coarse simulation for large networks, network timetable optimisation
Micro-Macro	Feasible train path/timetable optimisation

Table 2.1: Infrastructure models and their use cases.

calculations in the microscopic model are not changed. These include train route used, dwell times and supplements. However, in timetabling generation, a macroscopic model is often used which neglects the choice of train route causing infeasible solutions. In this case, a two-way conversion may be used where macroscopic solutions are checked for feasibility using a microscopic model. If the microscopic model finds the macroscopic solution infeasible additional constraints can be added to the macroscopic model before new solutions are generated.

See Schlechte et al. (2011) for a one-way micro-macro approach applied for optimisation of train path allocations, and Besinovic et al. (2015) for a two-way approach applied to timetable generation. In this thesis, a one-way micro-macro approach can be used with the mesoscopic model developed in chapter 7. See table 2.1 for an overview and comparison of infrastructure models.

2.3 Headway times and the blocking time model

One of the most important parameters for timetabling, simulation, and capacity planning is the minimum allowed headway as it directly affects the capacity available for train services.

In 1959 the idea of the blocking time model was conceived by Oskar Happel (Happel, 1959). This is now the most common method of modelling headways and is able to model modern advanced signalling systems (Pachl, 2002). As the name suggest, in this model the time a train ”blocks” a block section in the signalling system is modelled. As depicted in figure 2.2, a block is the track section between two consecutive signals or marker boards in a fixed block signalling system. In other terms, the blocking time is the time span where a given train has the exclusive right to a given block section. This also includes the time the block must be empty to ensure that the approaching train has enough time to brake if the block section is indeed not empty. As signal positions are necessary to model the

blocking time it naturally follows that a microscopic infrastructure model has to be used. The blocking time for a given train in a given block depends on (Pachl, 2008):

- Signal (and switch) set time
- Signal sight and reaction time
- Approach time (time to travel the braking distance)
- Running time through the block section
- Clearing time
- Release time

The signal set and release times are the time it takes for the signalling system to react on state changes, set routes and switches, and transmit the signal aspect to the driver via physical track side signals or through the driver machine interface in the driver's cabin (cab signalling). According to Klabes (2010), in a technical report from DB in 1992, this takes 27-120 seconds for mechanical systems, 12 seconds for electronic systems and 3 seconds multiplied with the number of switches for electro-mechanical systems.

The driver sight and reaction time is the time it takes from the time the driver can see and realise the signal aspect to the time braking can be initiated. The time depends on the operational rules. In Denmark, this time is set to 3 seconds for the driver to realise the signal, and 3 seconds reaction time from the signal has been realised until full braking is initiated (Banedanmark, 2013).

As the braking distance of trains is usually much longer than the sighting distance to the block signal, the signal has to be given in advance. This can either be done via preceding block signals, distant signals (which duplicate the block signal), or cab signalling. The braking capabilities of the rolling stock govern the minimum distance for which braking can be initiated and thus where the indication of the block signal aspect should be given. As the track side equipment usually is not aware of the braking capabilities of the train, the indicating distance has to be calculated with the onboard equipment and given to the driver via cab signals.

In the case that the train has no cab signalling and the track side equipment cannot distinguish between different types of rolling stock, a maximum braking distance is used for all trains. In Denmark, and many other European countries, this is the braking distance under normal circumstances with an acceleration of 0.6 m/s^2 plus technical reaction times (Banedanmark, 2013). If rolling stock cannot obtain this braking acceleration a

speed restriction is put on the rolling stock in question. Furthermore, speed restrictions may be put into force in extreme weather conditions to counter reduced friction.

The approach time is the time it takes for the train to travel the distance from the indication of the block signal aspect (where braking can be initiated) to the actual block signal. As mentioned above the indication may be given by physical signals or using cab signalling. With cab signalling it is possible to give a continuous update of the block signal state to the driver, while physical signals only allow for discrete updates. The way the indication is given significantly affects the duration of the approach time. E.g. with discrete updates every 2,000 meters (at every block signal) and a braking length of 3,000 meters of a particular rolling stock type the signal indication has to be given 4,000 meters in advance. For a continuously updated system the signal indication can be given instantly, i.e. 3,000 meters in advance. Thus a system with cab signalling (continuous update) provides a capacity benefit due to shorter blocking times given by the reduced approach times. The only exemption from this is the case where the braking distance matches the distance between the block signal and the indicating signal. However, this will rarely be the case.

An example of a signalling system with discrete update of the block signal aspect is the common European system ETCS level 1, where the driver is updated at each block section via cab signalling (blocks are physically marked by marker boards) (Theeg and Vlasenko, 2009). This system can be improved by using an inductive loop of a given length in front of the signal. This gives the system continuous update where the loop is placed. ETCS level 2 is an example of a system with continuous update where the block signal state is transmitted by radio to the driver (Theeg and Vlasenko, 2009). This system is now being implemented on the Danish long distance network.

The running time through the block section is the time it takes for the train to pass through the block section. This time may be modelled as described in appendix A, which is also the case for the approach time and the clearing time (described below).

The clearing time is the time it takes for the tail of the train to clear the block section and the following safety distance (overlap). The clearing time depends on the speed and length of the train as well as the length of the safety distance. In Denmark the safety distance is normally set to 150 meters corresponding to a braking distance from 40 km/h including reaction times (Banedanmark, 2013).

In figure 2.4, the blocking time is calculated for a train at two different block sections using the blocking time model. At *block a* the train traverses the block section with constant speed. At *block b* the train has a stop at the end of the previous block section which means that the approach time can be neglected as the train is already at a

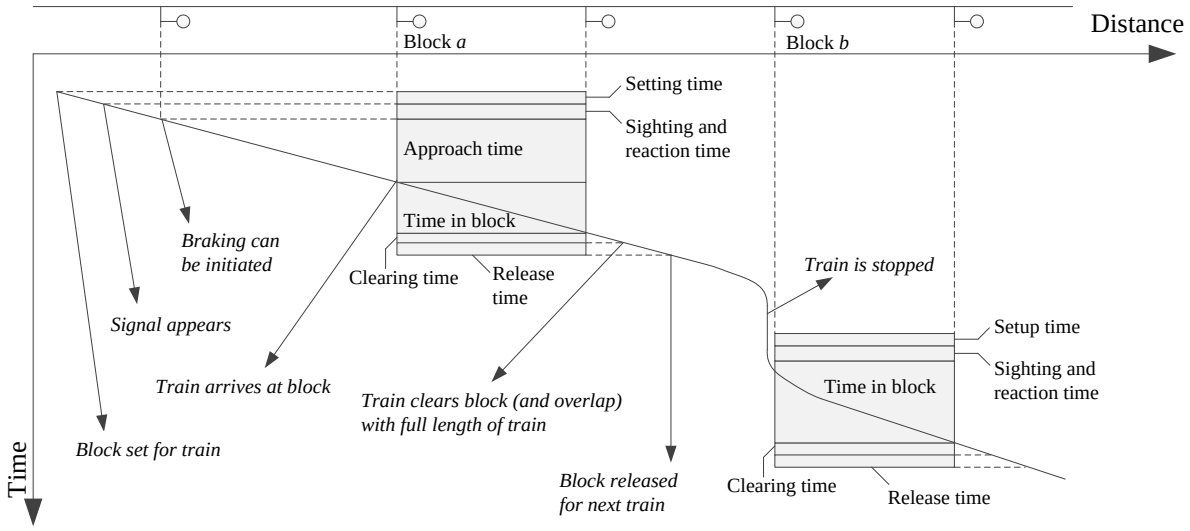


Figure 2.4: Blocking time for a train at two blocks. *Block a*, where the train travels through the block section at a constant speed, and *block b* where the train has a stop just in front of the block signal. Events are shown for *block a* only.

standstill (no braking distance). Although this reduces the blocking time, it also results in increased running time through the block section due to acceleration compared to *block a*. Furthermore, the dwell time of the train as well as the deceleration of the train will naturally increase the blocking time in the block where it has the stop. This increased blocking time in relation to stops is one of the main reasons that stations often constitute the capacity bottleneck in railway systems (one other reason being route conflicts).

To derive the minimum allowed line headway time between two trains (h_{ij}), as depicted in figure 2.5, the following equation can be used (Pachl, 2008):

$$h_{ij} = max(t_{be,1(k)} - t_{bb,1(k)}) \quad \forall k \in B \tag{2.1}$$

Where k is a block in the set of blocks B, $t_{bb,i(k)}$ is the start time of blocking block k for train i and $t_{be,i(k)}$ is the end of the blocking time for block k for train i. Eq. 2.1 thus gives the minimum separation time at the beginning of the considered line such that the second train can run freely without being hindered by the first train.

In addition to the fixed block signalling systems described here, there also exist moving block systems. By using such systems the minimum headway can be reduced significantly as the running time through the block is eliminated. Moving block systems are in use on closed networks with homogeneous rolling stock and are denoted CBTC (communication based train control). ETCS also contains a level 3 which is a moving block system, however this is not in used on any long-distance network due to challenges with train

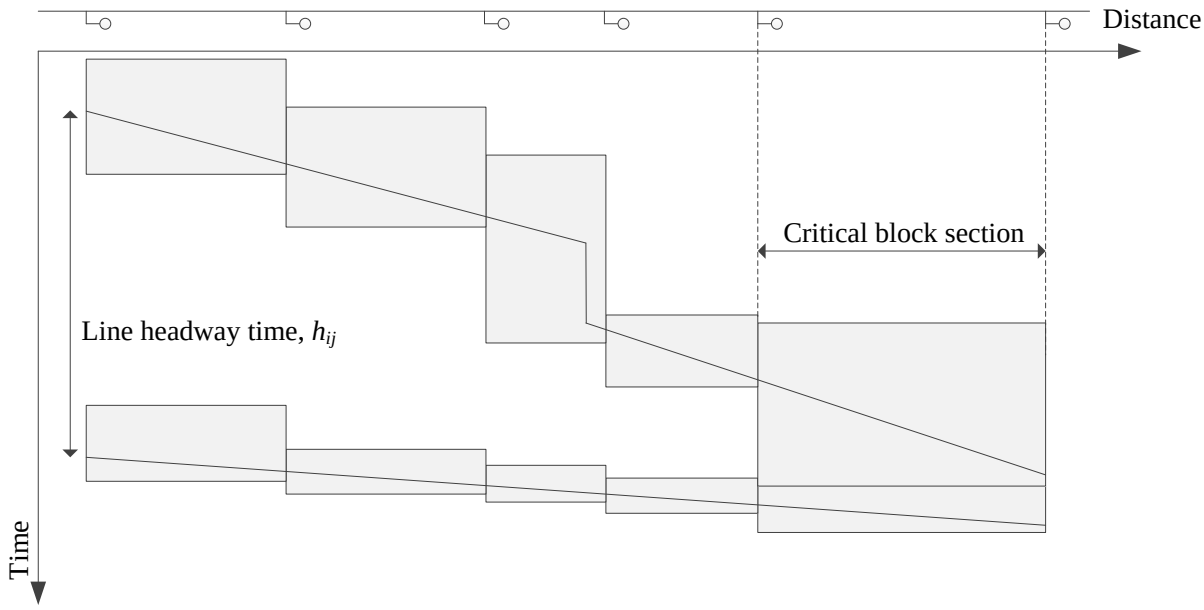


Figure 2.5: Minimum allowable headway time between two trains. The first train's movement authority is divided into two separate parts at the stop.

integrity checks of trains. A version of level 3 exists known as ETCS regional suitable for lines with homogeneous rolling stock.

2.4 Summary

In this chapter, planning and modelling of railways have been briefly introduced to give an overview of the terminology used for the thesis. Planning can be divided into three levels of planning: strategic, tactical, and operational. The methods described in this thesis is concerned with the strategic and early tactical planning, thus the planning horizon is long. Mainly, the processes of infrastructure planning and early timetables are suitable for the methods described in chapters 4 to 9. Modelling of railway infrastructure can be divided into microscopic, mesoscopic, and macroscopic models. As the names suggests, the level of detail differs from micro to macro, with microscopic models offering the highest precision. In this thesis, mesoscopic and microscopic models are used. In relation to the microscopic model, the blocking time model was introduced in this chapter, which is crucial to the calculation of headway times between trains and thus estimation of capacity, simulation, and timetable planning.

3 Service reliability in railways – Timetable robustness, stability and resilience

To maintain the trust of passengers in the published timetable and gain new passengers a reliably railway system is essential. This has gained more and more attention as the importance of service and travel time reliability has been recognised (van Oort, 2011; Cha, 2010). The increased research within this area has led to terms like robustness, stability and resilience of timetables being used more often.

In this chapter, an introduction to delays and delay reducing measures is given. This provides the basic terminology for a review on robustness, stability and resilience of timetables. Furthermore, the chapter is concluded with an introductory section on methods to evaluate robustness, including a short description of robustness indicators which is the main topic of part 1 (chapters 4 and 5).

3.1 Delays and supplements

In Europe, railway operations are organised and communicated by means of a timetable. However, operations are subject to process variations and external factors. Deviations from the deterministically planned timetable are therefore, inevitably, causing delays of passengers and freight to a greater or lesser extent. The objective in operations planning and real-time management is thus to reduce delay through proactive measures (buffer times and time supplements) and reactive measures (cancelling and/or re-routing of trains).

Delays can be categorised in different ways depending on how they arise. Throughout literature there is not a completely consistent definition. In this thesis, the following definitions are therefore used.

Disturbance, the natural variance in time duration of train processes causes (smaller) delays of trains. Examples of this are passenger numbers deviating from calculated values resulting in dwell time extensions, train driver behaviour causing running time devia-

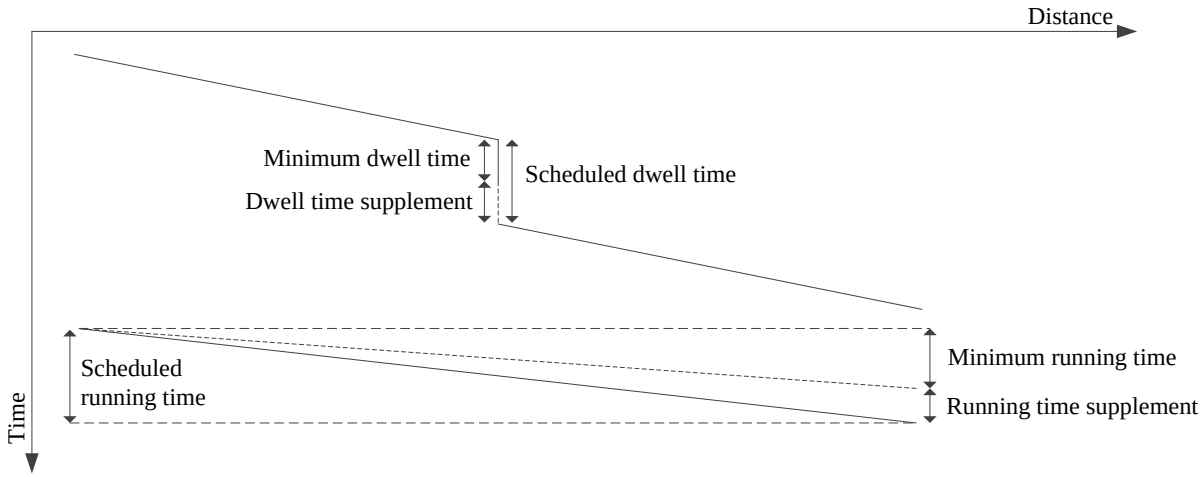


Figure 3.1: Running time supplement.

tions, weather conditions causing extended travel times and occasionally slower passenger exchange.

Disruptions cause larger delays, that typically arise from rolling stock, track, overhead line or signal/switch failures, or external factors like weather, strikes or accidents. As indicated by these examples, they tend to occur much more rarely than the smaller delays caused by process variance, however due to their nature they also tend to cause lengthy capacity reduction causing large delays.

Initial delay, also denoted primary delay, is the direct delay inflicted by a disturbance or disruption.

Consecutive delay, also denoted knock-on delay and at times secondary delay, is a delay imposed on a train caused by the delay of another train. The consecutive delay may arise due to track conflicts, transfer constraints, or dispatching decisions. If the network is highly utilised, a delay may propagate due to track conflicts caused by insufficient capacity.

To account for the smaller variances in running and dwell time a **supplement** is added to the minimum calculated running and dwell times. These are defined as running and dwell time supplements and illustrated graphically in figure 3.1. Alternative definitions include timetable slack, margins, or recovery times. When a smaller delay occurs, the supplements can be used to reduce or completely absorb the delays by speeding up the train or minimising the dwell time to the minimum.

When larger delays occur and the supplements added are not enough to absorb the delays, the delay will propagate from train to train causing consecutive delays. To reduce

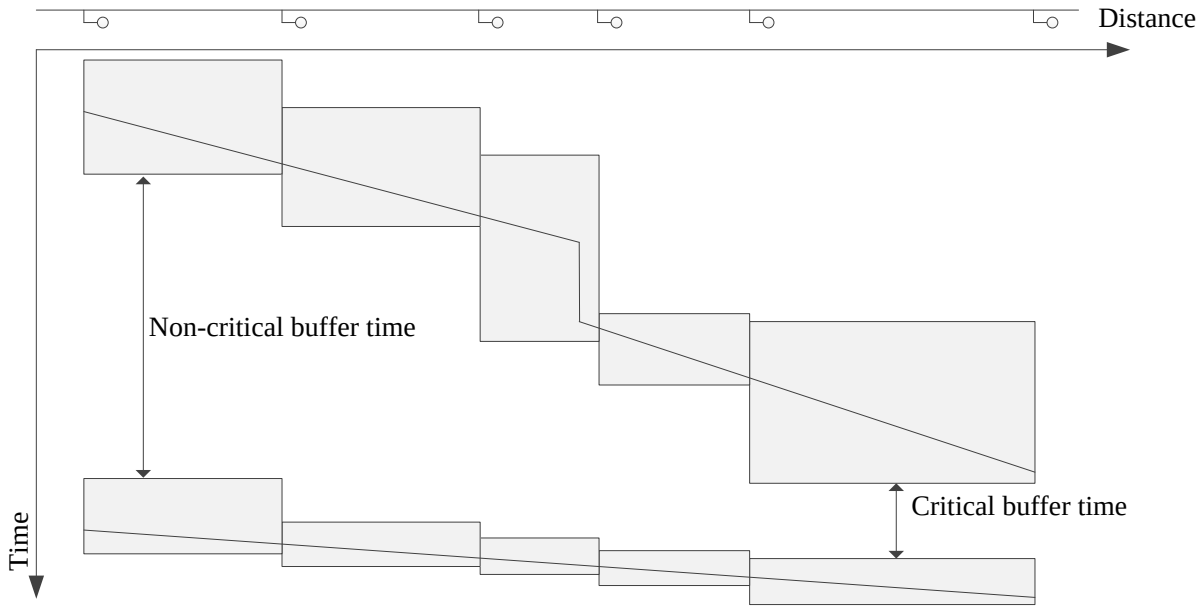


Figure 3.2: Buffer time between train paths. The first train's movement authority is divided into two separate parts at the stop

the risk of delay propagation, **buffer times** are placed between trains as shown in figure 3.2. Buffer times are occasionally also denoted slack. If the delay of a train exceeds the buffer time to the next train (in the same or opposite direction), the consecutive train will also suffer a delay due to a track conflict.

Both the amount of supplements and where they are added are crucial for the reliability of the timetable (Kroon et al., 2008a). In this context, Kroon et al. (2007) find that the running time supplements should be distributed primarily on the first part of a trip, to reduce or eliminate delay early to ensure (near) on-time performance on the remainder of the trip. A delay occurring late on a trip will have a lesser effect on the total trip and should therefore be assigned less supplement.

Caimi (2009) proposes the use of condensation and compensation zones to distribute supplements. A compensation zone is e.g. a line between two (larger) stations with spare capacity. In the compensation zone, the majority of the running (and dwell) time supplement is placed to allow trains to catch up on delay and arrive at time at larger stations, where the capacity is usually limited (condensation zones). This ensures a more efficient use of capacity as the risk of consecutive delays at stations is reduced and the number of trains handled (throughput) can be increased.

This concept is used in the Danish S-train network (Schittenhelm, 2011), where the capacity utilisation in the central part of the network is very high. The running time supplement is placed outside the central part of the network to maximise on-time trains at

arrival to the central part of the network where trains are then pushed through with little supplement to maximise the throughput of trains. This, however, has the disadvantage that trains that are not delayed run much slower than the allowed line speed outside the central part of the network, which results in longer scheduled travel times.

Regarding the amount of supplements, a supplement of 7% is usually used plus a supplement for larger stations in the Netherlands (Goverde, 2005). At the Danish infrastructure manager supplements in the range 3 to 13% are used (Schittenhelm, 2013). These supplements by the Danish infrastructure manager are larger when compared to the UIC's recommendations (Schittenhelm, 2013) which lie in the range 3 to 7% plus a distance-dependent supplement of 1 or 1.5 minutes per 100 kilometres (UIC, 2000).

Without considering the distribution of supplements and buffer times, a very reliable timetable can be achieved in theory simply by increasing time supplements and increasing buffer times by removing or reordering trains. However, the addition of excessive time supplements and low capacity consumption has a negative impact on travel times and the frequency of services. There is thus a delicate balance between supplements and fast travel times. Furthermore, experience has shown that excessive running time supplements may lead to reduced motivation of train staff and dispatchers to keep trains precisely on time (Schittenhelm, 2013). Thus in practice large running time supplements may increase the risk of delays compared to smaller supplements that keep the staff focused.

As described earlier, larger disturbances may also cause a reduction of capacity which must be dealt with by revising the timetable temporarily. To effectively change from the original timetable to a temporarily schedule, predefined contingency plans and real-time traffic management may be used. This reduces delay propagation by rescheduling and cancelling trains according to the current state of the system.

3.2 Definitions of robustness, stability, and resilience of timetables

As described in section 3.1, the amount and distribution of buffer times and time supplements are crucial for the reliability of the system together with effective transitions to contingency operation when larger delays and break-downs occur. In this sense a timetable (and system) may be described as stable, robust and resilient if it offers reliable service under all circumstances. These terms have distinct definitions in the literature with some variations especially regarding robustness. In the remainder of this section the different definitions are therefore reviewed.

Mattsson and Jenelius (2015) define stability generally as the ability of a system to return to equilibrium after a perturbation and the speed at which is does so. More specifically Goverde and Hansen (2013) state that a stable timetable can absorb initial delays by the use of time supplements without using active dispatching, while buffer times prevent or absorb consecutive delays. This is supplemented by Goverde (2005) who describes a stable timetable as one which prevents delays from circulating (propagating) in the network. There is a consensus in literature that timetable stability has a time/speed component. Thus a stable timetable can return to normal operation quickly after a disturbance (Vromans, 2005). Goverde (2008) states this explicitly as: "A system is globally stable if initial delays can settle in finite time". If only part of the network is considered, local stability can be measured as the ability to maintain the sum of output delay at or below the sum of input delay (Goverde, 2008). In relation to the speed at which the system settles, German rules for operation state that a 10 minute delay should be absorbed within 2 hours for a cyclic timetable (Pachl, 2002).

A common definition of robustness used by many authors is the ability of a timetable to absorb delays caused by smaller disturbances (Kroon et al., 2008a; Goverde, 2005; Salido et al., 2012; Andersson et al., 2013; Huerlimann et al., 2009; Vromans, 2005). Goverde and Hansen (2013) define these disturbances as caused by design errors, variance in parameters and changing operation conditions. If running time supplements are not enough to absorb initial delays, buffer times between trains must be able to prevent delay propagation (Kroon et al., 2008a; Andersson et al., 2013; Goverde and Hansen, 2013). Or as Vromans (2005) state: "When a railway system is not robust, small external influences cause large delays which propagate quickly [...]." Furthermore, Goverde (2005); Kroon et al. (2008a); Fischetti et al. (2009); Salido et al. (2012); Andersson et al. (2013), state in their definitions of robustness that the timetable should be able to cope with delay with little or no use of dispatching measures (changes). Thus the structure of the timetable must remain the same. As Fischetti et al. (2009) also state, major disruptions have to be handled in real-time rescheduling and are therefore not a part of the robustness definitions.

In relation to the definitions above there is no focus on the amount of supplement and buffer time which should be added to obtain a robust timetable. While larger supplements will yield a robust timetable, it will also cause a reduction in the level of service as described in section 3.1. As Cacchiani and Toth (2012) also describe, the duration and placement of buffer times in a timetable is a very important question that needs to be answered to guarantee a good trade-off between efficiency (fast travel times and high frequency) and robustness (delay resistance). Schöbel and Kratz (2009) describe this trade-off as "the price of robustness". Designing a robust timetable is therefore also

a matter of thorough analysis of the amount supplements and buffer times needed at critical locations without being excessive (Goverde, 2005).

Dewilde (2014) recognises this and extends the common robustness given above to the following: "A railway system that is robust against the daily occurring, small disturbances minimizes the real weighted travel time (RWTT) of the passengers". Thus the system should be utilised such that the average realised travel time is minimised. In this relation the amount and distribution of supplements and buffer times should be enough to cope with the frequently occurring process variation, but no more to allow for fast travel times when the services run according to schedule. E.g. Liebchen et al. (2010) find that it is possible to obtain a significant decrease in passenger delay with only a little increase in passenger travel times. Furthermore, Piester and Thorhauge (2010) find that a timetable supplement of 6.9-9.3% is optimal from a socio-economic perspective for Sydbanen in Denmark using simulation and a passenger delay model. The 12-13% originally planned by the Danish Transport Authority is therefore proved to be excessive. For more information on robustness with a passenger (efficiency) focus and on optimisation methods related to this definition, see Dewilde (2014) and Parbo (2015).

From the definitions described above two angles on robustness can thus be highlighted. One where robustness of the timetable is obtained with the use of supplements and buffer times without explicitly mentioning the amount of these. The other angle being an extended definition where an efficiency (no excessive use of supplements and buffer times) aspect is included.

Resilience of timetables relates to the timetables' ability to cope with larger disruptions through dispatching to reduce delay propagation. This is in contrast to stability and robustness which deal with delays without (or only light) dispatching measures as described above. Mattsson and Jenelius (2015) describe resilience as the capability of a system to persist when exposed to changes or shocks. Goverde and Hansen (2013) define this more precisely in railway terms as "the ability and flexibility of a timetable to handle larger disruptions by reducing or preventing secondary delays using dispatching". Thus a resilient timetable can easily and quickly be revised according to a new system state. In timetable planning, the resilient timetable is planned by maximising the effect of dispatching decisions to reduce delay propagation. Dispatching options available at the real-time management phase are e.g. re-ordering, re-routing, cancelling and skipping stops of trains. Furthermore, if plans for contingency are available, the resilient timetable should be able to quickly transform from (disrupted) normal operation to contingency operation.

From the literature review above, the following definitions are synthesised by this thesis:

- A **stable** timetable is a timetable that can settle from initial delay within a given time period without dispatching measures.
- **Robust** is the ability of a timetable to withstand smaller disturbances and variance in process times with little or no use of dispatching measures.
- A timetable that is **robust and efficient** can provide a service that minimises the realised travel times of passengers (and freight) (an extension of the robustness definition above).
- The **resilient** timetable can quickly make a transition into a new stable state when larger disruptions occur using dispatching measures.

In this thesis, the focus is primarily on obtaining stable and robust timetables, and the robust utilisation of railway capacity during the strategic and tactical planning phases. In this connection, robustness is defined as above in this thesis.

3.3 Robustness evaluation

Evaluation of the robustness of a timetable can be divided into two sets of methods. One is by measuring the performance of the timetable or alternatively by using indicators that, based on timetable characteristics, can give an indication of a timetable's robustness. Andersson et al. (2013) describe these two distinct sets of methods as ex-ante or ex-post analysis of the timetables' robustness. Where ex-ante analysis includes the use of indicators and ex-post analysis is the measurement of the performance of the timetable.

The performance of the timetable can either be measured based on the actual performance of the timetable or by the use of simulation. Typically the actual service provided compared to the announced service (the timetable) is given in the form of a punctuality measure and a reliability measure. These measures are key performance indicators (KPIs), and are the most commonly used to evaluate the performance of a contracted train operating company and its ability to operate on-time trains as set out in the contract. Thus these measures also provide insight into the stability, robustness, and resilience of a timetable.

The **punctuality** measure is the relative number of trains arriving/departing within a certain time threshold. This measure naturally depends on the threshold set out, but

also where it is measured. Is it arrival or departure at each station or only the terminal stations? This is quite important for the sake of measuring timetable performance. For a contracted train operator, an imprecise way of measuring punctuality may induce an operator to cheat. If only departure is measured the operator can put in extra dwell time supplements to reduce the delay of an arriving train, thus increasing the probability that trains will be punctual at departure. If the punctuality is only measured as the arrival at the terminal station, the operator can put in a significant amount of running time supplement at the last part of the train run to reduce possible delay. Thus a situation may arise where a train can be punctual at the terminal station, but not at any of the preceding stations.

The ideal way of measuring punctuality is to measure the punctuality of the content of the trains, that is the passengers or the freight. In this way, transfers are taken into account as well as how full the trains are. E.g a train may be punctual with a delay just within the punctuality threshold, but this minor delay may cause a missed transfer for some passengers which in the end are not punctual as they have to wait for the next train. However, in some cases a delayed train may also lead to passengers arriving before time. I.e. when transfer passengers are able to take a (delayed) train that they would otherwise miss if it was ontime. However, even due to this, passenger punctuality is generally lower than train punctuality as also shown by Landex (2008a) and Parbo (2015).

Measuring passenger punctuality is much more difficult than measuring train punctuality as passenger flows are needed. This can be done with smart card data (where available) and/or estimated with passenger route choice models, however this is not widely used. If passenger punctuality cannot be measured, the (train) punctuality should be measured at arrival at each station, preferably weighted by the amount of arriving passengers (or freight) for more precision.

The **reliability** measure is the relative number of trains actually operated. Thus a cancelled train will have a negative impact on reliability. In this context there is also the question of when the measurement is done. Is it when the full length of the train run has been cancelled or only part of it? The reliability can also be measured per stop yielding better detail. This is indeed relevant when dispatchers use the option of skipping stops in high frequency systems to reduce delay. As was the case with punctuality, the passengers (or freight) should be in focus. If this is indeed the case and punctuality is measured for passengers and freight, the reliability measure is not needed as e.g. a passenger will be (in most cases) delayed (not punctual) when a train is cancelled. In relation to simulation, the reliability measure is only relevant when larger disruptions are modelled and dispatching (with the possibility to cancel trains) is simulated.

In the following two sections simulation will be described as a measure of timetable robustness performance. Furthermore, a short introduction is made to robustness indicators which is the topic of chapters 4 and 5.

3.3.1 Simulation

Obtaining the actual performance of a proposed timetable can rarely be achieved. To test the expected performance of a timetable, simulation of a virtual representation of the railway system can therefore be used. In simulation the delay propagation in the timetable based on a set of input delays is modelled. From the amount of input delay absorbed by supplements and the extend of delay propagation, it is possible to derive the actual service offered and the deviation from the planned timetable (punctuality and reliability). Furthermore, it can by concluded whether the timetable is stable, robust, and resilient (if dispatching measures are simulated).

A simulation of railway operations may be done using either a microscopic, mesoscopic or macroscopic infrastructure model as introduced in section 2.2.2. For a precise simulation of the performance of a timetable, the microscopic model should be used. This is caused by the fact that trains might be slowed down due to delays. This in turn affects the headway time, which has to be re-estimated based on the signal blocks (a microscopic element) to obtain high precision. If meso- or macroscopic simulation models are used, this precision in headways cannot be achieved. Furthermore, the high abstraction in meso- and macroscopic models means that the effects of dispatching measures like re-routing cannot be properly captured. Therefore, it will often be necessary to use a microscopic model to derive if a timetable is resilient. In general, microscopic models should always be preferred for simulation purposes. However, if data is insufficient and/or to reduce computation time, macro- and mesoscopic simulation models can be useful or even necessary.

Given an input delay sampled from one or more probability distributions, the simulation algorithm has to figure out the following: can the delay be absorbed by supplements and will the delay propagate to (an)other train(s)? Based on the delay, rescheduling measures may have to be taken to reduce delay. This can either be done by a synchronous approach or an asynchronous scheduling approach.

In the asynchronous approach trains are rescheduled successively based on the priority of trains (Siefer, 2008; Jacobs, 2008; Nash and Huerlimann, 2004; Kaas, 1998b). Thus the algorithm seeks to reduce, eliminate or prevent delay of the train with the highest priority. In this context trains with low priority may be given even more delay. This approach corresponds to the natural course of timetabling where trains are timetabled

using a train hierarchy (Jacobs, 2008; Kaas, 1998b). The asynchronous approach has the advantage that new conflict-free timetables can be generated fast and dead-locks on single track lines can be avoided. This is mainly due to the fact that an asynchronous approach exploits the knowledge of the timetable and all the input delays. A disadvantage of the approach is the dependency on train priorities which may lead to large amounts of waiting time for lower prioritized trains, resulting in a low quality service (Kaas, 1998b). Furthermore, in real operation there is not a complete knowledge of input delays, which is a prerequisite for the asynchronous approach (at least within a certain time horizon).

On the contrary, synchronous simulation uses a traditional time-step model where trains are rescheduled simultaneously within a given discrete time step (Jacobs, 2008). The synchronous approach cannot foresee conflicts in advance as the asynchronous approach can. Only very obvious conflicts can be detected and thus deadlocks may occur. While the asynchronous approach reflects the timetabling process well, the synchronous approach reflects the actual operation better than the asynchronous approach. This is due to the fact that knowledge of future delays is not available, thus the synchronous approach yields a more realistic simulation of operation.

Two of the most popular simulation tools available are OpenTrack and RailSys. These systems are targeted towards the simulation of the railway operation based on a given timetable. Both systems use a microscopic representation of the infrastructure and a synchronous rescheduling approach to resolve conflicts that arise due to delays (Nash and Huerlimann, 2004; Bendfeldt et al., 2000). Other synchronous models are VISION, RailPlan, SIMONE, FALKO, TRANSIT, RAILSIM, RTC (Siefer, 2008), SCAN (Kaas, 1998b), and EGTRAIN (Quaglietta, 2011). While asynchronous models include ASDIS, BABSI and STRESI (Siefer, 2008; Jacobs, 2008). Some of these models also incorporate optimisation approaches.

As stated earlier in this section, a set of initial delays is given as an input to the simulation. Given the input delays, the delay propagation is estimated for the timetable, which sums up to the total delay (initial plus consecutive delays). In this context, an open question remains: what is the amount and distribution of input delay yielding a realistic simulation of the total delay? Two methods may be used to estimate initial delay distributions as described in the two following paragraphs.

First, a very common method is to calibrate the delay parameters based on observed data for a timetable that has been in operation. This is usually done by various ad hoc methods, based on performance measures and simple statistics, to adjust model delay parameters by trial and a error (Koutsopoulos and Wang, 2007). When the simulation results of the timetable (which has been in operation) are close to the actually observed

values, the model and input delays have been calibrated. A perfect calibration will in most cases be impossible to obtain as the calibration of a simulation model is a NP-complete problem (Hofmann, 2005).

The calibration obtained can then subsequently be used in new timetable scenarios. If the new timetable is very different from the timetable on which the calibration is based, the calibration will naturally lose some precision as preconditions for the calibration are changed. The microscopic simulation done in chapter 5 uses this approach.

The second approach to obtain initial delay parameters is to use some recommended parameters based on empirical data. E.g. based on some characteristics of a station, a certain probability distribution may be recommended based on passenger numbers (and flow) and the amount of signals, switches, etc. that may fail. Furthermore, details on the probability of failure of infrastructure elements like tracks and OCS as well as rolling stock can be used to estimate input delays.

In this thesis, a simulation model using a mesoscopic model has been developed to estimate critical buffer times (chapter 7). This simulation is based on an event network and thus uses a special synchronous approach where the time-step corresponds to the time between succeeding events. This is also generally denoted as a discrete event simulation. Furthermore, for obtaining input delays in this mesoscopic simulation model, general recommendations for probability distributions are used (the second approach described above). The (first) calibration approach cannot be used for the mesoscopic simulation model, as the preconditions for the calibration change significantly throughout the execution of the model.

As an alternative to simulation, max-plus algebra can be used to derive the stability or robustness of a timetable. Based on running times, minimum headway times, supplements, and initial disturbances, the timetable's ability to absorb these delays can be derived. As Goverde (2008) uses this method on a macroscopic representation of the infrastructure it has some of the same disadvantages as a macroscopic simulation in the form of reduced precision.

3.3.2 Indicators

As an alternative to simulation indicators can be used. The benefit of indicators compared to a simulation based approach is that they are very fast and usually requires much less data. This makes them suitable for early planning stages and optimisation models, where complex simulation models are too slow and cumbersome to use. Indicators for robustness

are based on data from one or more of the following sources: the infrastructure, (input) delays, and the timetable (structure).

Robustness indicators are the topic of the following two chapters. In chapter 4, existing indicators are reviewed and compared. Furthermore, a new indicator is suggested in chapter 4. In chapter 5, a case study is carried out where both indicators and microscopic simulation are used to study the ability of selected indicators to indicate robustness of a timetable. Thus ex-ante indicators are compared with the more precise (simulated) ex-post measures.

3.4 Summary

In this chapter definitions of robustness, stability, and resilience of railway systems have been reviewed. The review conducted is summarised in the following definitions:

- A **stable** timetable is a timetable that can settle from initial delay within a given time period without dispatching measures.
- **Robust** is the ability of a timetable to withstand smaller disturbances and variance in process times with little or no use of dispatching measures.
- A timetable that is **robust and efficient** can provide a service that minimises the realised travel times of passengers (and freight) (an extension of the robustness definition above).
- The **resilient** timetable can quickly make a transition into a new stable state when larger disruptions occur using dispatching measures.

The robustness measure given above serves as the definition throughout this thesis.

Furthermore, methods to assess robustness have been described and discussed in this chapter. Methods can be divided into two sets of methods, ex-ante and ex-post. Ex-post methods include measuring deviations from the timetable based on simulation or the actual timetable performance. Ex-ante methods include the use of robustness indicators to measure robustness of a timetable based on the its characteristics. Robustness indicators are computationally more simple and less data intensive than simulation and are therefore suitable for optimisation models and early planning phases.

Robustness indicators is the subject of the next two chapters, while a mesoscopic simulation model is used in chapter 7 to assess the robust consumption of capacity in a railway network based on a plan of operation.

Part I

Robustness indicators

4 Review and development of robustness indicators

This chapter is based on the following articles:

L. W. Jensen and A. Landex. Measuring Robustness of Timetables at Stations using a Probability Distribution. In *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis*, Copenhagen, 2013a

L. W. Jensen, A. Landex, and O. A. Nielsen. Evaluation of Robustness Indicators using Railway Operation Simulation. In C. Brebbia, N. Tomii, P. Tzieropoulos, and J. Mera, editors, *Computers in Railways XIV*, pages 329–339, Rome, 2014. WITpress. doi: 10.2495/CR140271

A. Landex and L. W. Jensen. Measures for track complexity and robustness of operation at stations. *Journal of Rail Transport Planning & Management*, 3(1):22–35, 2013

Robustness indicators can be used as an alternative, or supplement, to microscopic simulation when the robustness against delays has to evaluated for a railway system. A large number of indicators exist that directly or indirectly can be used to indicate the robustness of a railway system based on its characteristics. Robustness indicators that can be calculated fast are a good alternative in early planning/screening phases instead of simulation or to be used in objective functions in optimisation problems. This chapter contains a review of possible robustness indicators as well as new robustness indicators developed by this thesis.

The indicators reviewed can theoretically indicate the robustness of a timetable either partially or completely according to the definitions discussed in section 3.2. Specifically, the indicators reviewed have been selected based on one or more of the following criteria: ability to capture the size and distribution of initial delays, buffer times and time supplements.

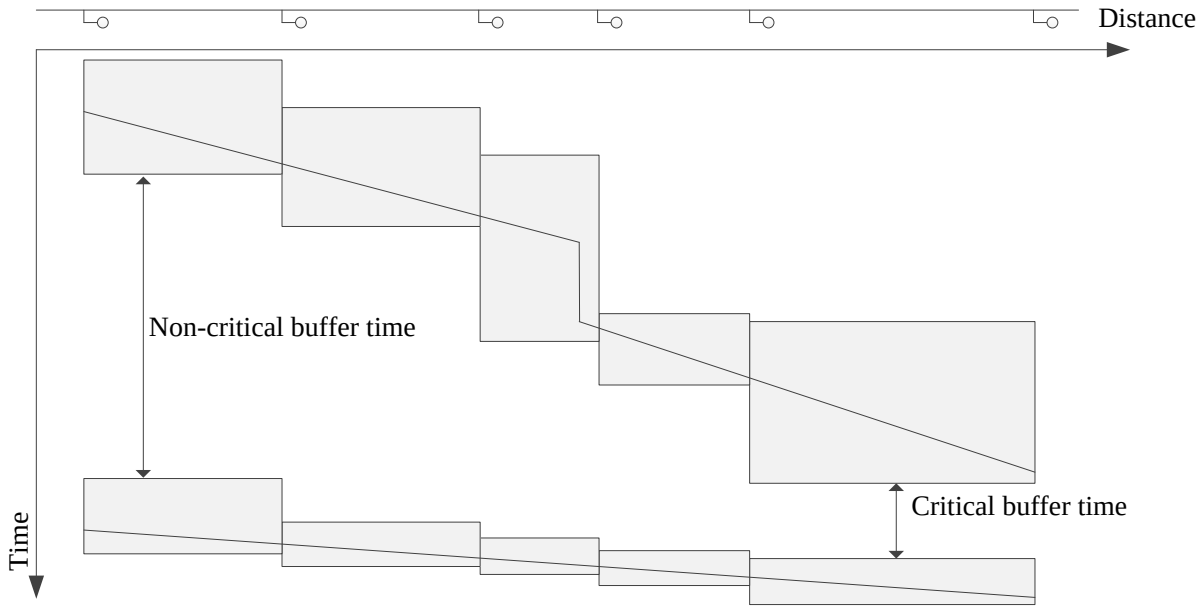


Figure 4.1: UIC 406 compression of a line section.

Initially in this chapter, robustness indicators based on buffer times, train heterogeneity, supplements, network locations with a high delay propagation risk, and travel time are reviewed (sections 4.1 to 4.5).

Subsequently, in section 4.6.1, a series of infrastructure complexities indices for stations are reviewed. These are the basis for a timetable complexity index developed by this thesis (section 4.6.2.5). This index makes it possible to estimate (indicate) the robustness of a specific timetable at stations (where delay propagation often occurs). Throughout section 4.6.1, the infrastructure and timetable complexities are applied and showcased on the Danish station Skanderborg. The complexity methods and results obtained for the case are discussed in section 4.6.3.

The chapter is summarised in section 4.7, where a theoretical comparison of the reviewed and developed indicators is also done.

4.1 Infrastructure occupation – UIC 406

The UIC 406 capacity method (UIC, 2004b) is an analytical method used to measure infrastructure occupation on railway lines. The method uses the approach of compressing train paths. The infrastructure occupation is expressed as the ratio between the completion time of the uncompressed and the compressed timetable. This ratio is thus an expression of the minimum buffer time between consecutive trains. This is easily observed

as the compression method eliminates the minimum buffer time between consecutive train paths as seen in figure 4.1. This relation can be expressed as in eq. 4.1.

$$\Sigma t_b^{min} = T - t_{compressed} \tag{4.1}$$

In this context it is important to note that the division into line sections is important. The longer the sections, the lesser information is contained in the ratio. This is especially important when the traffic is very heterogeneous. This can be seen by considering a fast train followed by a slow train between two overtaking/junction stations. The smallest buffer time will generally occur at the beginning of the section and the largest buffer time at the end of the section. Using the UIC 406 method on such a section will only reveal the minimum buffer time, thus neglecting the larger buffer times at the end of the section. The UIC 406 method is described in further detail in section 6.4.

4.2 Heterogeneity indices

Heterogeneity indices can be used to measure the distribution of trains on line sections and at stations. These indices can be used to indirectly indicate the robustness of a timetable by measuring the spread of buffer times. Heterogeneity indices are not suitable for lines with scheduled bidirectional operation (single track lines for example).

Vromans (2005) proposed a number of heterogeneity indices that are able to measure the distribution and heterogeneity of trains over a given period. Two of these are SSHR (sum of shortest headway reciprocals) and SAHR (sum of arrival headway reciprocals). SSHR is based on the shortest headway times, h_i^-, between consecutive trains on a track section. The second measure, SAHR, is based on the arrival headway times between trains, h_i^A. This measure is proposed as arrival heterogeneity seems to be more important as fast trains catch up on slow trains at the end of sections (Vromans, 2005). The research by Vromans (2005) suggests that reducing the value of SSHR and SAHR, and thus increasing homogeneity, increases robustness as the risk of delay propagation from smaller initial delays is decreased.

$$SSHR = \sum_{i=1}^{n} \frac{1}{h_i^-} \tag{4.2}$$

$$SAHR = \sum_{i=1}^{n} \frac{1}{h_i^A} \tag{4.3}$$

Both SAHR and SSHR are based on the absolute headway times. Thus they capture both heterogeneity of headways and the absolute value of headway times. However, in some cases the objective may be to only capture the heterogeneity of headways independent of the size of these. Based on Landex and Jensen (2013), the following headway heterogeneity index, HH, is therefore suggested by this thesis for a cyclic timetable, where h_i is the headway time at arrival or departure:

$$HH = 1 - \left(\sum_{i=1}^{n-1} \min \left(\frac{h_i}{h_{i+1}}; \frac{h_{i+1}}{h_i} \right) + \min \left(\frac{h_1}{h_n}; \frac{h_n}{h_1} \right) \right) \cdot \frac{1}{n} \tag{4.4}$$

This indicator also allows for the possibility to compare timetables (or sections) where the number of trains differ.

4.3 Distribution of running time supplement

Kroon et al. (2007) use a stochastic optimisation approach to distribute the running time supplement along a railway line to optimise delay reduction based on a delay sampled from an exponential distribution. A measure denoted as the weighted average distance (WAD) is used to express the center of the running time supplement on a scale from 0 to 1. A value of 0.5 means that the same amount of running time supplement is placed on the first and last part of the journey. Thus if the running time supplement is distributed proportionally along the line, the WAD will be 0.5. A value less than 0.5 indicates that more supplement is placed on the first part of the journey and a value above 0.5 that the majority of supplement is placed on the second part of the journey. As described in section 3.1, the running time supplement should primarily be placed in the start to the middle of the journey. Thus WAD should be slightly below 0.5. In the computational experiments by Kroon et al. (2007), values of 0.32 and 0.492 are found to be optimal in two cases with different amounts of delay.

The weighted average distance (WAD) can be expressed as:

$$WAD = \sum_{tr=1}^{N_{tr}} \frac{2 \cdot tr - 1}{2 \cdot N_{tr}} \cdot t'_{tr} \tag{4.5}$$

Where tr is a part of a train run between two timetable points, N_{tr} is the total number of trips and t'_{tr} is the running time supplement on trip tr.

4.4 Point measures

In a railway network, certain points inherently impose a risk of reduced robustness caused by the high risk of delay propagation due to train dependencies. These points are typically at stations. The following two subsections (4.4.1 and 4.4.2) present two methods concerning these points defined as fix and critical points respectively.

4.4.1 Train path fix points and risk profiles

Schittenhelm and Landex (2013) propose to use fix points to measure the complexity of a train path. Fix points are points in the timetable where a train path is dependent on another train path and thus it implies a risk of delay propagation. Fix points are closely related to the complexity indices described in section 4.6. Schittenhelm and Landex (2013) list the following fix points for train paths:

- Overtaking stations where the train path is scheduled to overtake (or be overtaken by) another train (path)
- Crossing stations on single track railway lines where a train path is scheduled to cross a train in the opposite direction
- Transition stations where the number of tracks available to the train path is reduced, and either a crossing is scheduled or a specified train order has to be kept
- Stations where a train path catches up with a slower path according to the timetable

In addition to this, Schittenhelm and Landex (2013), also define two additional categories of fix points, one for transfer stations where transfers between trains have to be kept, and one with fix points for crew and rolling stock.

Given the fix points, Schittenhelm and Landex (2013) propose the following robustness indicators for a train path:

- Number of fix points per train path (or group of train paths) as measured per section (according to UIC 406 divisions, see section 6.4) or per path kilometre
- Risk profile for a train path (or group of train paths) giving the amount of running and dwell time supplements between fix points

The (relative) number of fix points gives an indication of the complexity of the train path and thus the risk of the train path suffering delay propagation.

The risk profiles can be used as an indicator of the timetable's ability to absorb delays on the different sections to avoid delay propagation at fix points.

4.4.2 Critical points

Andersson et al. (2013) propose the RCP measure (Robustness in Critical Points). This indicator focuses solely on critical points in space and time where the risk of delay propagation from a train to the next is high. Andersson et al. (2013) define the following two situations as being critical points:

1. Where a train enters a line after a train already operating on the line
2. An overtaking station where a train is planned to be overtaken by another train

For each of the critical points identified, the RCP indicator is calculated as (Andersson et al., 2013):

$$RCP_p = L_p + F_p + H_p \tag{4.6}$$

Where L_p is the running time supplement available before the location of the critical point for the first train in the train combination vulnerable for delay propagation, F_p is running time supplement available for the second train after the location of the critical point, and H_p is the buffer time between the two trains at the critical point location. Thus RCP_p is the total amount of time available to absorb and/or prevent delay propagation at critical points.

Andersson et al. (2015) use the RCP indicator to improve robustness of the timetable on part of the main line between Stockholm and Malmø in Sweden. Specifically, a MILP model is used to maximise the value of RCP (the robustness) by increasing running time supplements and buffer times for each critical point. The increased supplement is obtained by redistributing running time supplements from other parts of the line while minimising the changes done to the original timetable. The study by Andersson et al. (2015) shows that additional robustness can be obtained, using the RCP indicator, by redistributing supplements and buffer times.

4.5 Robustness indicator including passenger travel time

As mentioned in section 3.2, Dewilde et al. (2014) includes the aspect of efficiency by measuring robustness as the weighted travel time extension (WTTE). This is calculated as the relative difference between the total weighted real travel time and the nominal travel time of passengers:

$$WTTE = \frac{\text{total weighted real travel time} - \text{nominal travel time}}{\text{nominal travel time}} \tag{4.7}$$

The nominal travel time is the minimum necessary time to make a journey without running and dwell time supplements, but including necessary transfers (Dewilde et al., 2014). The total weighted real travel time is the actual travel time of passengers, where weights are given according to each of the travel time components. E.g. arrival delays and missed transfers are given a weight 3 times higher than the minimum running time (Dewilde, 2014). The lower the value of WTTE, the more robust (and efficient) the timetable is. This can be seen as the total weighted real travel time having to be as close as possible to the nominal travel time. In practice, this means that the optimal balance between supplements and delays suffered must be found to minimise the expected travel time of passengers.

As the total weighted real travel time is based on delays, these has to be obtained either by simulation or from actual recorded delays. Thus WTTE is not a robustness indicator in line with the other indicators described in this chapter, but rather a performance indicator for a (simulated) timetable. As such the WTTE measure is only relevant for simulation, timetable analysis and simulation-optimisation of timetables. On the contrary, the other indicators mentioned in this chapter can be used by a manual approach and across a wider range of optimisation problems as these indicators can be expressed analytically.

4.6 Complexity indices

As mentioned in section 4.4, stations are often points of delay propagation due to train dependencies. This is a natural consequence of conflicting train routes in the switch zones. The risk of delay propagation is highly dependent on the infrastructure layout (complexity) of the station as well as the number of trains using the station. A station's robustness in relation to conflicts can be estimated by a complexity index for the given

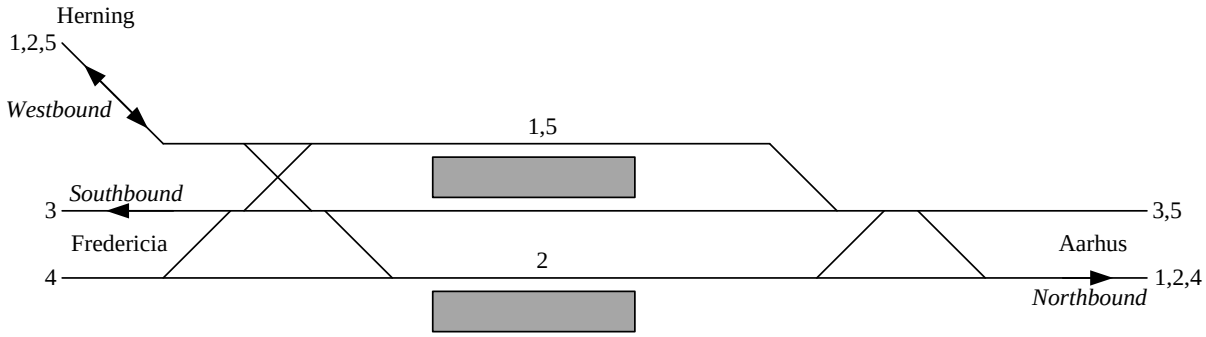


Figure 4.2: Main routes used in normal operation at Skanderborg station.

station. A high complexity index implies a low robustness due to a higher risk of delay propagation caused by conflicting routes.

In the following sections, existing complexity indicators are reviewed. Subsequently, a new complexity indicator is developed by this thesis that is able to capture the robustness of the operation at a given station.

The junction station Skanderborg in Denmark is used as an example throughout these sections on complexity indices. At Skanderborg station, the single track line from Herning converges with the main line from Copenhagen and Fredericia. The single track line is characterised by regional trains running between Herning and Aarhus. The main line is characterised by IC and ICE traffic stopping at Skanderborg and through-going regional trains.

The infrastructure of Skanderborg is shown in figure 4.2 where the main routes used in normal operation are enumerated. Two routes, 1 and 2, have been included for trains from Herning towards Aarhus. Route 2 is used when trains to and from Herning need to cross at Skanderborg, if this is not the case route 1 is used. Other routes are not considered although they may be used in real-time rescheduling when a train is delayed.

4.6.1 Methods for calculation of infrastructure complexity indices

Three methods exist for the calculation of complexity indices for a station based on the route locking table and thereby the infrastructure of the station. The three methods are presented in the following three subsections and provide increasingly detailed results.

4.6.1.1 Infrastructure complexity based on infrastructure

The simplest method to calculate the complexity is purely based on the infrastructure layout of the station and therefore does not take the timetable or plan of operation into account. This complexity index is calculated as the ratio between the sum of conflicting routes and all routes according to Kaas (1998b); Landex and Jensen (2013); Lindner (2011); Pachl (2002). Routes that cannot be set after each other are discarded in the calculation as these route combinations are not feasible. Usually this is only the case with routes to and from dead-end tracks. Infrastructure complexity can be expressed as in eq. 4.8

$$\varphi_n = \sum_{i,j \in R} c_{ij} \cdot \frac{1}{N} \tag{4.8}$$

In eq. 4.8 N is the number of feasible route combinations and c_{ij} is a binary variable indicating if there is a conflict between route i and j or not (an infeasible route combination is regarded as non-conflicting).

The number of conflicting routes and the total sum of routes can be found by using a route locking table. In table 4.1 the route locking table (main routes only) for Skanderborg is shown. All route combinations are feasible as there are no routes that cannot be set after each other.

Table 4.1: Route locking table for Skanderborg.

Route: 2nd \ 1st	1	2	3	4	5
1	O	O	X	C	X
2	O	O	X	C	X
3	X	X	O	–	D
4	C	C	–	O	–
5	X	X	D	–	O

The following notation holds for table 4.1:

- : No conflict ($c_{ij} = 0$)

O: Overlapping (or partially overlapping – diverge to later converge again) routes ($c_{ij} = 1$)

C: Converging routes ($c_{ij} = 1$)

D: Diverging routes ($c_{ij} = 1$)

X: Crossing routes ($c_{ij} = 1$)

Using eq. 4.8, the complexity index, based on the infrastructure, φ_n, of Skanderborg can be calculated as 21/25 (0.84), as there are 21 conflicting route combinations and 25 combinations in total. The complexity index calculated is rather high (maximum is 1) as routes 1 and 2 are in conflict with each of (all) the other four routes.

4.6.1.2 Infrastructure complexity based on probability of conflict

The simple way of calculating the complexity index for a station, presented in section 4.6.1.1, does not take the number of trains using each route into account. According to Kaas (1998b); Landex and Jensen (2013); Pachl (2002) the complexity of the station when taking the infrastructure and number of trains using each route into consideration can be calculated by the means of the probability of a conflicting route combination:

$$p_{ij} = \frac{n_i \cdot n_j}{n^2} \tag{4.9}$$

Where n_i is the number of trains using route i, n_j is the number of trains using route j and n is the total number of trains using the station. Again only feasible route combinations should be included in the calculation.

For Skanderborg the probabilities for each feasible route combination have been calculated using eq. 4.9. The number of trains using each route is derived from the 2012 timetable on a weekday between 15:00 and 16:00. Table 2 shows the results.

Table 4.2: Probability of conflict at Skanderborg.

Route 1st / 2nd	1 ($n_1 = 1$)	2 ($n_2 = 1$)	3 ($n_3 = 3$)	4 ($n_4 = 3$)	5 ($n_5 = 2$)
1 ($n_1 = 1$)	0.01	0.01	0.03	0.03	0.02
2 ($n_2 = 1$)	0.01	0.01	0.03	0.03	0.02
3 ($n_3 = 3$)	0.03	0.03	0.09	0.09	0.06
4 ($n_4 = 3$)	0.03	0.03	0.09	0.09	0.06
5 ($n_5 = 2$)	0.02	0.02	0.06	0.06	0.04

The complexity of the station can then be calculated as the ratio between the sum of probabilities where there is a conflicting route combination and the sum of probabilities of route combinations, only considering feasible route combinations:

$$\varphi_p = \sum_{(i,j)\in R} (p_{ij} \cdot c_{ij}) \cdot \frac{1}{\sum\limits_{(i,j)\in R} p_{ij}} \tag{4.10}$$

This is similar to the method presented in section 4.6.1.1, with the difference that the relative use of the routes is taken into account.

Using eq. 4.10 the complexity, φ_p, for Skanderborg can be calculated as 0.70, where the sum of probabilities for all route combinations is 1 and the sum of probabilities for conflicting route combinations is 0.70.

By taking the relative use of routes into consideration, the calculated complexity of the station is reduced from 0.84 to 0.70. This is because more trains use routes 3 and 4 than route 5 and especially routes 1 and 2. The method presented in this section gives a more realistic result of the complexity of the station than the method presented in section 4.6.1.1. However, as the method is based on the relative use of the different routes, the complexity will not increase even though the number of trains is doubled. In real life this would cause a major increase in complexity of the operation at the station.

4.6.1.3 Infrastructure complexity based on headways

The third method expands on the previous two methods by taking minimum headway times for a given time period into account. In this way it is possible to evaluate both the infrastructure and the operation at the station, but also taking the total number of trains into consideration.

Estimated minimum headway times (in seconds) for the different route combinations are shown in table 4.3 for Skanderborg. It is important to stress that these headway times are estimated as real life minimum headway times are difficult to obtain (Schittenhelm and Landex, 2013).

Table 4.3: Estimated headway times in seconds at Skanderborg.

Route: 2nd \ 1st	1	2	3	4	5
1	180	180	240	180	240
2	180	180	240	180	240
3	240	240	180	–	180
4	180	180	–	180	–
5	240	240	180	–	180

According to Kaas (1998b); Landex and Jensen (2013); Potthoff (1962), the occupation time of each train route can be calculated as:

$$f_{ij} = \frac{p_{ij} \cdot c_{ij}}{\varphi_p} \cdot h_{ij} \tag{4.11}$$

Using eq. 4.11 the occupation times of the train routes at Skanderborg have been calculated in seconds and are shown in table 4.4.

Table 4.4: Expected (weighted average) occupation times in seconds at Skanderborg.

Route 2nd \ 1st	1	2	3	4	5
1	2.6	2.6	10.3	7.7	6.9
2	2.6	2.6	10.3	7.7	6.9
3	10.3	10.3	23.1	–	15.4
4	7.7	7.7	–	23.1	–
5	6.9	6.9	15.4	–	10.3

The total time the station is occupied, t_o, can be calculated using:

$$t_o = \varphi_p \cdot n \cdot \sum_{(i,j)\in S} f_{ij} \tag{4.12}$$

Taking the time period examined into consideration the complexity can then, according to Potthoff (1962), be expressed as:

$$\varphi_W = \frac{t_o}{T} \tag{4.13}$$

Using eq. 4.12 & 4.13 the complexity, φ_W, of Skanderborg is 0.38 with a total occupation time, t_o, of 23 minutes and a time period, T, of 60 minutes. The complexity of operation calculated for Skanderborg, φ_W, is lower than both φ_n and φ_p calculated in sections 4.6.1.1 and 4.6.1.2. This is a consequence of the relatively low minimum headway times at the station. If the minimum headway times or number of trains were increased, the complexity, φ_W, would also increase.

4.6.2 Timetable complexity

In this section, two methods are presented to calculate the complexity of a station where the timetable is also considered and not only the plan of operation. Thereby, increasing precision of the complexity index compared to the methods presented in section 4.6.1.

The methods presented in this section rely on delay data and a probability distribution that can describe train delays. A probability distribution that can describe train delays given a set of train delays is therefore presented in section 4.6.2.1 together with empirical train delay data for Skanderborg Station.

The two methods are used on three different timetable alternatives at Skanderborg station to illustrate the potential of the two methods. The three timetable alternatives are presented in section 4.6.2.3. The first method is presented in section 4.6.2.4 and is based on a threshold, where a high risk conflict is detected if the buffer time of a train combination is less than this threshold. The method is used on three timetable alternatives at Skanderborg with a threshold equal to the mean delay of the first train in a train combination.

The second method, developed by this thesis and presented in section 4.6.2.5, is based on a probability distribution that fits the delay data for the station and a given train combination. The method is used on the three different timetable alternatives at Skanderborg with the probability distribution and delay data from section 4.6.2.2.

4.6.2.1 Delays and Probability Distributions for Delays

A probability distribution fitting the train delays of a given station is an essential part of the method presented in section 4.6.2.5. Finding a suitable probability distribution for delay data depends on whether the distribution has to fit all delays or only non-negative delays. This thesis proposes that only non-negative delays are considered for the probability distribution as negative delays are not delays in a strictly sense.

According to Schwanhäußer (1974) the non-negative arrival delays at a station can be described by an exponential distribution. This is, in most cases, confirmed by Goverde et al. (2001) in a study where delay data at Eindhoven station in the Netherlands were examined. Furthermore, Goverde et al. (2001) also find that departure delays can be fitted well by an exponential distribution.

These findings are somewhat supported by the work of Yuan (2006, 2007), who finds that the Weibull distribution generally gives the best fit of non-negative arrival and departure delays on the analysed data at the Hague HS station in the Netherlands. In many cases Yuan (2006, 2007) finds that the Weibull distribution can be fitted to the delay data with a shape parameter around 0.5-1.2. This supports the theory that the exponential distribution is generally a good fit for non-negative delays as the Weibull distribution with a shape parameter of 1 is exactly the exponential distribution. However, these findings also show that the Weibull distribution can give a better overall fit than the exponential distribution. Furthermore, in some cases Yuan (2006, 2007) also finds that the gamma, beta, or log-normal distribution gives a better fit than the Weibull distribution.

In addition to this, several proposals have been made for more flexible distributions that can be used to achieve better fits on delay data, e.g. an extended exponential distribution as proposed by Büker and Seybold (2012).

On the basis of the above, this thesis suggests that the exponential distribution is used as a generic distribution to describe the non-negative arrival and departure delays at stations when calculating the timetable complexity index using the method presented in section 4.6.2.5. However, any cumulative distribution function can be used.

The cumulative distribution function (cdf) for the exponential distribution is given as:

$$F(v;\lambda) = 1 - e^{-\lambda \cdot x} \tag{4.14}$$

The rate parameter, λ, for a given set of delay data that is exponentially distributed, can be found by:

$$\lambda = \frac{1}{E[X]} \tag{4.15}$$

Where the expected value of the exponentially distributed random variable X is equal to the arithmetic mean of the given delay data that is exponentially distributed.

4.6.2.2 Delay Data for Skanderborg Station

Delay data has been analysed for Skanderborg station with the purpose of estimating λ for the exponential distribution and the mean delay to be used in the calculation of the complexity index of Skanderborg station in sections 4.6.2.4 and 4.6.2.5. Data is provided by Rail Net Denmark's RDS system (punctuality and operations statistical system). The system contains records of delays at 330 measurement points. At 166 of these measurement points, data is automatically collected from the CTC system when a track circuit becomes occupied, with a precision of one second. For the remaining measurement points the delay is reported manually, typically with a precision of one minute, significantly reducing the quality of the delay data (Richter, 2012; Schittenhelm and Landex, 2013). For a station the system reports the arrival and departure delay for stopping trains or, in case of a through-going train, the delay when passing through.

The data for Skanderborg consist of train delay data, recorded manually, for the 2012 timetable (December 11, 2011 to December 8, 2012). To analyse consistent data, weekends and holidays, days with track work and the summer period from 23/06/12 to 12/08/12 have been excluded from the data. Furthermore, it has been chosen only to analyse trains in the afternoon peak from 15:00 to 16:00. The analyses have been done on two data sets; one with both non-negative and negative delays and one with only non-negative delays. The data set with all delays is included in the analysis only to represent the number of on time trains within a threshold of 3 and 5 minutes.

Table 4.5: Statistical overview over arrival delays at Skanderborg between 15:00 and 16:00.

Arrival		Non-negative delays			All delays	
Train	n	Median [s]	Mean ($1/\lambda$) [s]	Std. dev. [s]	< 3 min	< 5 min
IC144 (S)	201	60	65	79	0.90	0.97
IC145 (N)	190	10	68	181	0.91	0.96
L45 (N)	191	60	167	267	0.71	0.83
L46 (S)	204	60	78	130	0.88	0.95
RA5352 (N)	163	30	96	232	0.87	0.95
RA5357 (W)	153	0	26	68	0.96	0.98
RX5350 (N)	147	19	66	115	0.90	0.96
RX5355 (W)	159	0	27	58	0.95	0.99
RR3345 (N)	101	0	47	118	0.95	0.97
RR3348 (S)	89	78	136	188	0.75	0.90

Table 4.6: Statistical overview over departure delays at Skanderborg between 15:00 and 16:00.

Departure		Non-negative delays			All delays	
Train	n	Median [s]	Mean ($1/\lambda$) [s]	Std. dev. [s]	< 3 min	< 5 min
IC144 (S)	209	60	71	89	0.89	0.97
IC145 (N)	208	20	72	181	0.90	0.95
L45 (N)	207	60	164	267	0.70	0.82
L46 (S)	210	60	91	143	0.86	0.94
RA5352 (N)	164	0	64	207	0.92	0.96
RA5357 (W)	164	59	70	84	0.83	0.98
RX5350 (N)	163	0	61	110	0.90	0.96
RX5355 (W)	164	0	33	69	0.93	0.98

In the data set with non-negative delays some outliers have been excluded. An outlier is a data observation that differs significantly from the bulk of the data. Outliers have been identified in the data for IC144, IC145, L45, RR3345 and RA5357.

Table 4.5 shows a summarizing table of arrival delays at Skanderborg station. It can be seen that more than 90% of all trains, except InterCity Express 45 (L45), arrives within the Danish punctuality threshold of 4 minutes and 59 seconds for regional and long-distance trains (Richter, 2012). Furthermore, it can be seen that the northbound trains have a higher average delay than the south and westbound trains. This is probably caused by the fact that the northbound trains are further from their starting stations than the south and westbound trains.

Also, considering train direction, it can be seen from table 4.5 that the regional trains to and from Western Jutland (RA & RX) are generally less delayed than the InterCity

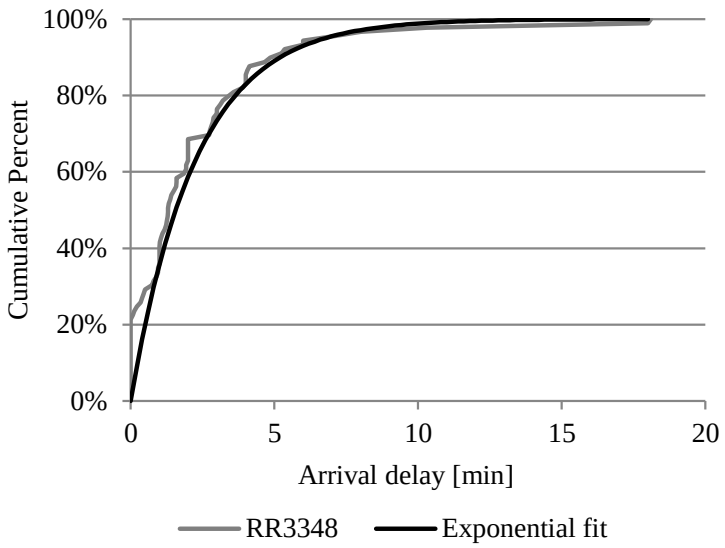


Figure 4.3: RR3348 arrival delay data and exponential cumulative distribution function with $\lambda = 0.44$.

trains and especially the InterCity Express trains. This is most likely due to the fact that the InterCity trains operate over a longer distance passing more bottlenecks and are therefore more likely to experience delays, even though they have a high priority.

The regional trains running between Esbjerg and Aarhus (RR) do not fit the pattern described which is probably due to unplanned changes of rolling stock.

Lastly, it is also seen that the median in many cases is not that close to the mean. This is mainly caused by the precision of one minute in most of the delay data.

Table 4.6 shows the departure delay for the trains departing from Skanderborg between 15:00 and 16:00. As expected the results are similar to the arrival delay data in table 4.5. However, it can be seen that RA5352 has a lower average departure delay than arrival delay. The reason for this can be found in the long dwell time of five minutes, which can be reduced when the train arrives delayed, thereby reducing the delay at departure. From tables 4.5 and 4.6 it can also be seen that RA5357 has a higher average departure delay than arrival delay. This can be explained by the fact that the RA trains cross at Skanderborg, while the RX trains do not. The westbound RA train therefore has to wait for the northbound train on the single track line towards Herning, illustrating a network effect that can cause consecutive delays.

For RR3348 the mean arrival delay is 2.26 minutes (see table 4.5). Using eq. 4.15, λ is calculated as 0.44 by the assumption that the delay data is exponentially distributed. Figure 4.3 shows this assumed fit as the cumulative delay data of RR3348 and the ex-

ponential cumulative distribution function with $\lambda = 0.44$. It is seen from the figure that an exponential distribution with $\lambda = 0.44$ yields a fairly good representation of the delay data of RR3348.

4.6.2.3 Timetables

Three timetables for Skanderborg station are analysed. All three have the same number of trains and same route usage, which means that the infrastructure complexity indices will be the same for all three. The three timetables are:

- The 2012 timetable on a weekday between 15:00 and 16:00
- Alternative 1 where small changes have been made to the 2012 timetable to increase the smallest buffer times
- Alternative 2 where changes have been made to the 2012 timetable to achieve small or no buffer times

Network effects of the timetable changes have not been considered in the creation of alternatives which may deem them infeasible. For the calculation of probabilities in section 4.6.2.5 the exponential distribution is used as proposed in section 4.6.2.1. For the method in section 4.6.2.4 the mean delay of the first train in a train combination is used as the threshold.

The rate parameter, λ, has been estimated using eq. 4.15 and the mean of non-negative delays listed in tables 4.5 and 4.6. As the λ-values (and the mean delay) are based on actual delay for the 2012 timetable good precision is achieved. However, for the two alternative timetables no delay data exist. Therefore, the assumption is to use the same mean and λ-values as the 2012 timetable in sections 4.6.2.4 and 4.6.2.5, with a loss of precision. To reduce this loss of precision, arrival and departure times for the trains to and from Herning are kept the same in all three timetables. The choice of fixing these trains rather than the IC, ICE and RE trains running on the main line, is based on the fact that the line to and from Herning is single track. A small change in arrival and departure times for trains to and from Herning at Skanderborg may result in rescheduling of crossings on the line to Herning. This may significantly change the delay characteristics and thereby the mean and λ-values for the trains on the single track.

The impact on the result and the λ-values of trains running on the double track line to and from Fredericia (see figure 4.2) is considered to be smaller. The choice of using the same mean and λ-values can only be recommended when relatively small changes are

made to the timetable. If larger timetable changes are to be assessed by these methods it is recommended that a set of more generic values are used, as preconditions are significantly changed.

Table 4.7: Timetable alternatives.

Timetable:		2012		Alt. 1		Alt. 2		λ	
Route (i)	Train (j)	Arr	Dep	Arr	Dep	Arr	Dep	Arr	Dep
1	RE N (1)	3	6	3	6	3	6	0.91	0.98
2	RE N (1)	31	36	31	36	31	36	0.63	0.94
3	ICE S (1)	40	41	40	41	3.5	4.5	0.77	0.66
3	IC S (2)	15	16	20	21	33	34	0.86	0.80
3	RE S (3)	47.5	-	47.5	-	40	-	0.44	0.44
4	ICE N (1)	26.5	27.5	53	54	28	29	0.29	0.29
4	IC N (2)	57	58	23	24	56	57	0.50	0.50
4	RE N (3)	17.5	-	17.5	-	9.5	-	0.94	0.94
5	RE W (1)	8.5	9.5	8.5	9.5	8.5	9.5	2.22	1.82
5	RE W (2)	28.5	31	28.5	31	28.5	31	2.31	0.82

Table 4.7 shows the departure and arrival times of the three timetables. Furthermore, the table also shows the estimated λ-values (inverse mean) for the different trains arriving and departing at Skanderborg. The timetables span 60 minutes.

4.6.2.4 Timetable complexity using a buffer time threshold

On the basis of the methods presented in section 4.6.1 Schittenhelm and Landex (2013) have developed a method to calculate the timetable complexity index at a station where a cyclic timetable is taken into account. The calculation procedure in this method is:

1. For each train, calculate the buffer time between the incumbent train and the train immediately preceding it on the same track (O, D or C conflict)

2. For each train, calculate the buffer time between the incumbent train and the train using a crossing route (X conflict) immediately before the incumbent train for each crossing route

3. If the buffer time for each train combination is below a predefined threshold, a high risk conflict has been identified

4. The timetable complexity index (φ_h) can then be calculated as the ratio between high risk conflicts (n_h) and the total number of potential (and feasible) high risk conflicts (n_{ph}):

$$\varphi_h = \frac{n_h}{n_{ph}} \tag{4.16}$$

Schittenhelm and Landex (2013) propose the following three ways of defining the threshold: a simple estimate (e.g. 3 minutes), using the mean delay of the first train in a train combination, or by a percentile in the delay data (of the first train in a train combination).

In the following, the method is used on the three timetables presented in section 4.6.2.3. The threshold parameter is chosen as the mean non-negative delay (estimated from delay data) of the first train in a train combination. Arrival and departure times together with inverse mean delay, λ, is listed in table 4.7 for the three different timetables. The calculation of high risk conflicts is shown in table 4.8.

The table essentially shows the calculation of the complexity index. For each train the buffer time between the incumbent train and the train immediately preceding it on the same track (O, D or C conflict) is calculated. Furthermore, the buffer time between the incumbent train and the train using a crossing route (X conflict) immediately before is calculated for each crossing route. This gives a potential of up to three high risk conflicts (O, D or C conflicts and the number of crossing routes) for a train at Skanderborg as seen from table 4.8. Furthermore, the threshold is shown in table 4.8 as the mean delay of the first train in the train combination. If the buffer time is below this threshold, a high risk conflict has been detected (marked bold).

From table 4.8, and using eq. 4.16, the complexity for the 2012 timetable is 1/44 (0.023), for timetable alternative 1 it is 1/44 (0.023) and for timetable alternative 2 it is 3/44 (0.068). The method is thus capable of somewhat capturing the large amount of small buffer times in alternative 2. However, the small improvements to increase the smallest buffer times in alternative 1 are not reflected in the complexity index calculated with this method. This is because the 2012 timetable only has one high risk conflict, which is also the case with timetable alternative 1. The small improvements made in timetable alternative 1 is not enough to eliminate this single high risk conflict thus resulting in the same conflict index.

4.6.2.5 Timetable Complexity using a Probability Distribution

This thesis suggests an alternative to the method presented in section 4.6.2.4. The suggested method provides more precision due to the use of a continuous probability distribution. Essentially, the timetable complexity index is calculated in the following steps:

Table 4.8: Calculation of complexity index for the three timetables at Skanderborg.

Timetable	2012			Alternative 1			Alternative 2		
Train (route i, train j)	Conflicting trains Buffer time [min] Threshold [min]			Conflicting trains Buffer time [min] Threshold [min]			Conflicting trains Buffer time [min] Threshold [min]		
1,1	4,2	3,31	5,2	4,1	3,3	5,2	4,2	3,3	5,2
	5	4.5	30.5	9	14.5	30.5	6	22	30.5
	1.2	2.3	0.4	2.7	2.3	0.4	1.2	2.3	0.4
2,1	4,1	3,2	**5,2**	4,2	3,2	**5,2**	4,1	3,1	**5,2**
	5.5	11	**0**	9	6	**0**	4	22.5	**0**
	2.7	1.2	**0.4**	1.2	1.2	**0.4**	2.7	1.5	**0.4**
3,1	5,2	1,1	2,1	5,2	1,1	2,1	3,3	1,1	2,1
	8.5	30	6	8.5	30	6	20.5	53.5	29.5
	0.4	1.0	1.6	0.4	1.0	1.6	2.3	1.0	1.6
3,2	3,3	1,1	2,1	5,1	1,1	2,1	5,2	1,1	**2,1**
	24.5	5	41	8.5	10	46	1.5	23	**0**
	2.3	1.0	1.6	0.5	1.0	1.6	0.4	1.0	**1.6**
3,3	3,1	1,1	2,1	3,1	1,1	2,1	3,2	1,1	2,1
	5	38	13	4.5	37.5	12.5	4	30	5
	1.3	1.0	1.6	1.3	1.0	1.6	1.1	1.0	1.6
4,1	4,3			2,1			4,3		
	6			15			15.5		
	0.8			1.1			0.8		
4,2	2,1			2,1			2,1		
	19			45			18		
	1.1			1.1			1.1		
4,3	1,1			1,1			**1,1**		
	8.5			8.5			**0.5**		
	1.0			1.0			**1.0**		
5,1	3,3	1,1		3,3	1,1		3,1	1,1	
	18	1.5		18	1.5		2	1.5	
	2.3	1.1		2.3	1.1		1.3	1.1	
5,2	3,2	1,1		3,3	1,1		5,1	1,1	
	10.5	21.5		5.5	21.5		17	21.5	
	1.1	1.1		1.1	1.1		0.5	1.1	
n_h		1			1			3	
n_{ph}		44			44			44	
φ_h		0.023			0.023			0.068	

1. For each train, the probability of delaying other trains that are using a conflicting route, within a time period, is calculated
2. The maximum theoretical probability of delays is calculated for each train
3. The complexity is then the sum of all values found in step 1 divided by the sum of all values found in step 2

For overlapping and diverging route combinations the train order cannot be changed and the second route will therefore suffer a consecutive delay if the first route is delayed more than the buffer time. For crossing and converging routes the train order can be changed and a delay will therefore only arise if the delayed train arrives when the track is occupied by the other train. However, in heterogeneous operation, the train order is crucial for the capacity of the system and a dispatcher may therefore choose not to change the train order from the planned one as it may cause even more delays.

The calculation of timetable complexity, φ_d, can be represented mathematically as:

$$\varphi_d = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{(t^b_{jk},h^m_{jk},q_{jk})\in S_{ij}} \bar{P}(t^b_{jk};h^m_{jk};q_{jk})}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{n_i}(n^s_i\cdot((1-\bar{F}_{ij}(T-h^m_{jj}))+\sum\limits_{(h^m_{jk})\in S_{ij}\setminus h^m_{jj}}(\bar{F}_{ij}(\sum\limits_{l=2}^{k}h^m_{jl})-\bar{F}_{ij}(T-h^m_{jj}))))} \tag{4.17}$$

Where:

i: 1st train route of a route combination

j: A train using route i

k: A train using a route that is in conflict with route i (j is included with index = 1)

l: Incremental counter for k

N: Number of routes

n_i: Number of trains using route i

n^s_i: The number of routes, in conflict with route i, that can be set at the same time. Can be identified by the use of a diagram of non-conflicting routes as proposed by Pachl (2002); Potthoff (1962).

T: Time period

h^m_{jk}: The minimum headway time between train j and k

q_{jk}: The type of conflict between train j and k (O, D, X or C)

The tail distribution of a probability distribution, $\bar{F}_{ij}(x)$, that yields a good representation of the expected delay of train j using route i as a function of x (the buffer time) is given by:

$$\bar{F}_{ij}(x) = 1 - F_{ij}(x) \tag{4.18}$$

Where $F_{ij}(x)$ is the cumulative distribution function of a probability distribution. The tail distribution yields a probability of 1 if there is no buffer time, depending on type of conflict, and decreasing probability as the buffer time increases.

S_{ij}, is a set of buffer times, t^b_{jk}, minimum headways, h^m_{jk}, and type of conflict, q_{jk}, for each train, k, that is in conflict with train j using route i:

$$S_{ij} = \{(t^b_{j1}, h^m_{j1}, q_{j1}), ..., (t^b_{jk}, h^m_{jk}, q_{jk}), ..., (t^b_{jm}, h^m_{jm}, q_{jm})\} \tag{4.19}$$

The timetable is assumed cyclic and therefore the corresponding train of the incumbent train in the next time period, j, is also included in S_{ij} with a buffer time of:

$$t^b_{jj} = T - h^m_{jj}. \tag{4.20}$$

Alternatively, the set S_{ij} may be defined similarly to Schittenhelm and Landex (2013). I.e. this thesis then defines S_{ij} as the entries $(t^b_{jk}, h^m_{jk}, q_{jk})$ where k are the trains that use a route in conflict with route i. However, for each route in conflict with i, only the train k closest in time to train j is considered. These combinations are the most significant for the results as they contain the train combinations with the lowest buffer times from each combination of conflicting routes.

The buffer time, t^b_{jk}, is calculated as the buffer time between train j and k taking the time period, T, into account:

$$t^b_{jk} = \begin{cases} (arr_k - arr_j) - h^m_{jk}, & \text{if } arr_j < arr_k; \\ (arr_k - arr_j) - h^m_{jk} + T, & \text{otherwise.} \end{cases} \tag{4.21}$$

Where arr_k is the arrival time of train k at the conflict point and arr_j is the arrival of train j at the conflict point. In addition to the above expression it assumed that the timetable is conflict free when planned which implies that t^b_{jk} is non-negative.

In eq. 4.17, $P(t^b_{jk}; h^m_{jk}; q_{jk})$ is a conditional tail distribution, which is necessary to introduce, as the probability of a conflict depends on the conflict type:

Table 4.9: Calculation of complexity index for train 1 using route 1, train 1 using route 4, and train 2 using route 5.

i	j	k (route, train)	t^m_{kj}	h^m_{jk}	q_{jk}	P
1	1	1 (1,1)	57.0	3	O	3.78E-28
		2 (2,1)	25.0	3	O	9.39E-13
		3 (3,1)	28.3	4	X	3.70E-13
		4 (3,2)	3.3	4	X	3.66E-02
		5 (3,3)	36.1	4	X	1.37E-16
		6 (4,1)	18.5	3	C	6.96E-09
		7 (4,2)	49.0	3	C	2.69E-22
		8 (4,3)	8.0	3	C	2.89E-04
		9 (5,1)	1.5	4	X	1.88E-01
		10 (5,2)	21.5	4	C	4.48E-11
4	1	1 (4,1)	57.0	3	O	1.76E-69
		2 (4,2)	27.5	3	O	6.74E-34
		3 (4,3)	48.0	3	O	1.26E-58
		4 (1,1)	35.5	3	C	7.72E-43
		5 (2,1)	1.5	3	C	1.55E-02
5	2	1 (5,2)	57.0	3	O	1.18E-11
		2 (5,1)	37.0	3	O	8.06E-08
		3 (1,1)	30.5	4	X	1.18E-06
		4 (2,1)	55.0	4	X	1.73E-28
		5 (3,1)	8.5	3	D	2.35E-02
		6 (3,2)	43.5	3	D	4.58E-09
		7 (3,3)	16.3	3	D	7.50E-04

$$\bar{P}(t^b_{jk}; h_{jk}; q_{jk}) = \begin{cases} \bar{F}_{ij}(t^b_{jk}), & \text{if } q_{jk} = O \text{ OR } q_{jk} = D; \\ \bar{F}_{ij}(t^b_{jk}) - \bar{F}_{ij}(t^b_{jk} + h^m_{jk}), & \text{otherwise.} \end{cases} \tag{4.22}$$

For the denominator of eq. 4.17 the sum of maximum theoretical probabilities for each train is calculated. The maximum theoretical probability for train 1 using route 4 (ICE N) is:

$$n^s_4 \cdot ((1 - \bar{F}_{4,1}(T - h^m_{1,k})) + \sum_{h^m_{1,k} \in S_{4,1} \setminus h^m_{1,1}} (\bar{F}_{4,1}(\sum_{l=1})^k h^m_{1,l}) - \bar{F}_{4,1}(T - h^m_{1,k})) \tag{4.23}$$

$$\Leftrightarrow 1 \cdot ((1 - \bar{F}_{4,1}(57)) + (\bar{F}_{4,1}(3) - \bar{F}_{4,1}(57)) + (\bar{F}_{4,1}(6) - \bar{F}_{4,1}(57)) + (\bar{F}_{4,1}(9) - \bar{F}_{4,1}(57))) = 1 \tag{4.24}$$

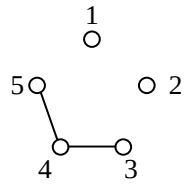


Figure 4.4: Diagram of non-conflicting routes at Skanderborg.

For train 1 using route 4, n_i^s is 1 as there are two conflicting routes, 1 and 2, however these routes cannot be set at the same time. This can easily be seen by constructing a diagram of non-conflicting routes as proposed by Lindner (2011); Pachl (2002); Potthoff (1962), see figure 4.4

In the diagram in figure 4.4 every route is represented by a node. Two nodes are connected if the two routes are non-conflicting. In this way, it is possible to see which routes can be set at the same time and how many routes that can be set at the same time. From figure 3, it can be seen that neither route 1 nor route 2 can be set simultaneous with any of the other routes. Route 4, however, can be set at the same time as routes 5 or 3. This means that n_i^s for routes 1 or 2 is 2 as both routes are in conflict with routes 3, 4 and 5, but two of these three routes can be set simultaneous.

Using the method described in this section on the three different timetables at Skanderborg station, the following complexity indices, φ_d, have been found:

- 0.115 for the 2012 timetable,
- 0.107 for timetable alternative 1,
- and 0.418 for timetable alternative 2.

The complexity indices reflect that some minor alterations have been made to increase the smallest buffer times in timetable alternative 1 compared to the 2012 timetable. By reducing the number of small buffer times the complexity is reduced. The results also show that timetable alternative 2 is much more prone to consecutive delays than both the 2012 timetable and alternative 1 due to a large amount of small buffer times.

4.6.3 Discussion and comparison of the complexities

Five methods have been presented in this section that can be used to indicate the robustness at stations. The three methods presented in section 4.6.1 are focused on the infrastructure of the station and should be used when planning the infrastructure and/or plan of operation, while one of the methods in sections 4.6.2.4 and 4.6.2.1 should be used

Table 4.10: Complexity indices for Skanderborg station and the three timetable alternatives.

Complexity method	2012 Timetable	Timetable alt. 1	Timetable alt. 2
Infra. - Track layout (φ_n)	0.84	0.84	0.84
Infra. - Probability (φ_p)	0.70	0.70	0.70
Infra. - Headway time (φ_W)	0.38	0.38	0.38
Timetable - Buffer threshold (φ_h)	0.02	0.02	0.07
Timetable - Delay probability (φ_d)	0.12	0.11	0.42

when timetable variants are assessed or the infrastructure is evaluated on the basis of one single timetable.

Table 4.10 summarises the results of the complexity calculations for Skanderborg station by complexity method and timetable alternative.

It can be seen from table 4.10 that the complexity decreases from φ_n to φ_p to φ_W. In the case of φ_p this is due to the fact that more trains use the routes that have fewer conflicts with the other routes. The further reduction in complexity, calculated by considering headway times, φ_W, is caused by relatively low headways. If these were to be higher the calculated complexity, φ_W, would be higher than φ_p. Infrastructure complexity indices calculated using the three methods presented in section 4.6.1 (φ_n, φ_p and φ_W) cannot be directly compared with the timetable complexity indices calculated using the two methods presented in section 4.6.2 (φ_h and φ_d). This is caused by the fact that the calculations are done fundamentally different.

Comparing the results for φ_h and φ_d and the methods presented in sections 4.6.2.4 and 4.6.2.1 it can be concluded that the method developed by this thesis using delay probabilities (section 4.6.2.1) is able to capture even small changes while the threshold method (section 4.6.2.4) is not able to do this. This is a natural cause of the nature of the two methods where the method based on delay probabilities is based on a continuous function while the method based on a threshold is based on a binary decision variable (the threshold).

The method presented in section 4.6.2.1 could be revised to also take probabilities of consecutive delays into account as noted in the definition of S_{ij}. However, this would not give the exact same results as the considered train combinations are not the same.

Specifically, the method presented in section 4.6.2.4 only considers trains using conflicting routes immediately before the incumbent train, for each train. While in the method presented in section 4.6.2.1 all conflicting train combinations are considered for each train. This gives considerably more calculations in the method presented in section 4.6.2.1. On the contrary, the conflicting trains do not have to be ordered for each train,

which has to be done with the method presented in section 4.6.2.4. Yet in most cases, the results will not be significantly different from each other as the most important train combinations (with the lowest buffer times) are always considered.

The two methods presented in sections 4.6.2.4 and 4.6.2.1 have been used with empirical delay data to calculate the complexity indices. This is an advantage when calculating the complexity index for a real timetable where delay data is available. However, when new timetable variants have to be assessed no delay data is available and existing delay data therefore have to be used. This is not a problem if the deviation from the timetable variant to the timetable, from which the delay data is used, is small. If there is a large difference the expected delay must be estimated.

The methods in section 4.6.2 have been used on three different timetables where timetable alternative 1 has been found to have the lowest complexity (see table 4.10). This is a local minimum of the three timetables at Skanderborg station and timetable alternative 1 is therefore not necessarily the least complex (and most robust) timetable when the whole network is taken into account. Thus complexity calculations have to be done for several (vulnerable) stations in the whole network to get a reasonable image of a timetable's robustness compared to that of another timetable.

4.7 Summary and theoretical comparison

In this chapter, indicators for robustness of railway systems have been reviewed. Table 4.11 summarises the reviewed indicators. The table shows the reviewed indicators' ability to capture size and distribution of initial delays, buffer times, and time supplements either directly or indirectly (marked in parenthesis). Furthermore, the applicability of the indicators on stations, lines or both is defined in table 4.11. The last column in the table indicates the input needed to calculate the robustness indicators.

Besides simulation (and WTTE, the weighted travel time extension indicator, which is evaluated by simulation) none of the reviewed indicators are able to capture initial delays, buffer times, and time supplements at the same time. The UIC 406 method can indirectly be used to indicate the amount of buffer times available through compression of the timetable. However, no indications of the distribution in time of such buffer times are given. For this purpose, headway heterogeneity measures can be used to indicate the distribution of headways in time at start/end of line sections, and thus the distribution of buffer times.

The distribution of time supplements may be measured by the WAD (weighted average distance) indicator, which measures the center of the amount of supplement added to a

	Initial delays		Time supp.		Buffer times		PTT	Applicable	Input
Indicator	Size	Dist(s)	Size	Dist(s)	Size	Dist(t)			
UIC 406					(•)			L	TT
WAD				•				L	TT
Heterogenity						•		L & S	TT
Inf. complexity					(•)			S	I (P)
TT complexity	•	•			•	•		L & S	TT D
Risk profiles			•	•				L & S	I P
RCP_p			•	(•)	•	(•)		L & S	TT
WTTE	•	•	•	•	•	•	•	A	TT D
Simulation	•	•	•	•	•	•	(•)	A	TT D

Table 4.11: The ability of robustness indicators to measure size and distribution, over distance (s) or time (t), of initial delays, buffer times or time supplements. PTT = passenger travel time. L = lines, S = stations, A = aggregated. I = infrastructure, TT = timetable (including I), P = plan of operation, D = delays

train run on a section. WAD is limited to a value between 0 and 1, where Kroon et al. (2007) suggest that the value should be slightly below 0.5, which indicates that the time supplements should be placed mainly on the first part of the train run considered.

Certain stations in a railway network pose a high risk of delay propagation due to train interdependencies. Such points in the network are identified as critical points or fix points in literature. In relation to these points, the amount (and distribution) of time supplements and buffer times is crucial to reduce or eliminate the risk of delay propagating between trains. A train path risk profile indicates the amount of fix points per train path and/or the timetable supplement available between fix points. The RCP_p is an indicator for the Robustness in Critical Points which is the sum of the time supplements before and after the critical point as well as the buffer time at the critical point.

The RCP_p measure requires the infrastructure and a timetable as an input, while the risk profiles (and fix points) only require the infrastructure and a plan of operation. However, the approach of fix points and train path risk profiles may be extended to include the minimum buffer times to the following train path, if the timetable is available (similarly to RCP_p).

In relation to stations that pose a risk of delay propagation due to route conflicts, three methods have been reviewed which deal with infrastructure complexity at stations with different levels of detail. These three methods can be used to indicate the robustness against delay propagation, and thus indirectly the buffer time, for different infrastructure layouts. The method described in section 4.6.1.1 is solely based on the infrastructure as an input. By adding the number of trains using each route to the calculation (section 4.6.1.2) a more accurate complexity result of the station and the plan of operation is given.

However, the total number of trains is not taken into consideration. This is achieved by taking headway times into account as presented in section 4.6.1.3. Thus the precision of the complexity result, for the station and the plan of operation at the station, is increased.

None of the three infrastructure complexity methods reviewed, take the exact timetable into account. This is introduced by two timetable complexity methods presented in sections 4.6.2.4 and 4.6.2.1, where the latter has been developed by this thesis. Based on the timetable, delay data, and a probability distribution the complexity of the timetable based on the infrastructure is calculated. This can be used to indicate the robustness of the timetable as the calculation is based on the actual buffer times between all train combinations at the station as well as initial delays.

The infrastructure complexities are focused on the infrastructure of the station and should be used when planning the infrastructure and/or plan of operation, while one of the methods in sections 4.6.2.4 and 4.6.2.1 should be used when timetable variants are assessed or the infrastructure is evaluated on the basis of one single timetable.

The five complexity methods have been applied on Skanderborg station in Denmark. The results show that the three methods for infrastructure complexity provide increasing detail correlated with the amount of input given. The timetable complexity index developed by this thesis (section 4.6.2.1) is able to capture changes in buffer times between three different timetable alternatives at Skanderborg station, thus extending the existing indicators.

From a planning phase perspective, infrastructure complexities are suited for strategic planning, fix points and UIC 406 for strategic to tactical planning (if the train order is known). While the remaining indicators reviewed are only suited for the (early) tactical level (timetabling) as a timetable is required. All indicators, except the complexity indices, can be used at the macroscopic to the microscopic level.

5 Evaluation of robustness indicators

This chapter is based on the following article:

L. W. Jensen, A. Landex, and O. A. Nielsen. Evaluation of Robustness Indicators using Railway Operation Simulation. In C. Brebbia, N. Tomii, P. Tzieropoulos, and J. Mera, editors, *Computers in Railways XIV*, pages 329–339, Rome, 2014. WITpress. doi: 10.2495/CR140271

While the semantics of a microscopic simulation are well-known, the semantics of robustness indicators are typically not. This makes it hard to use the robustness indicators for other purposes than comparison of alternatives as recommendations on absolute values are typically not given. Furthermore, not all robustness indicators may necessarily indicate robustness of infrastructure plans and/or timetables as good in practice as in theory (as reviewed in chapter 4).

This chapter therefore evaluates a selection of the robustness indicators described in chapter 4 by comparing indicator results with a microscopic simulation (using RailSys (Bendfeldt et al., 2000)) on the North West line in Denmark. This shows how well indicators perform compared to a detailed simulation and the indicators ability to capture robustness and possibly the semantics of indicators.

This chapter is organised as follows. Section 5.1 describes the methodology of this evaluation while the results are given and discussed in sections 5.2.1 and 5.2.2. Conclusions and a summary of the study are given in section 5.3.

5.1 Methodology

For the case study conducted in this chapter, the following indicators have been selected to represent the different parameters that affect robustness (size and distribution of initial delay, time supplements, and buffer times):

- UIC 406 (size of buffer times)
- Heterogeneity indices (distribution of buffer times)
- Fix points and risk profiles (size and distribution of running time and dwell time supplement)
- Infrastructure complexities (risk of delay propagation – buffer times)
- Timetable complexity (size and distribution of buffer times and initial delays)

The indicators are described in detail in sections 4.1, 4.2, 4.4.1 and 4.6.

The 80 kilometre long North West line in Denmark is used as a case. This line has been extended in 2014 from single to double track between Lejre (Lj) and Vipperød (Pe) (see figure 5.1). An increase of the line speed from 120 to 160 km/h is also part of the project. The capacity increase achieved is used to improve the service in the 2020 timetable (see figure 5.2). This yields the following four scenarios that together with the possibility to isolate the line from the network makes it a good case for this analysis.

1. 2010 timetable (with loko-hauled trains) on the old infrastructure
2. 2010 timetable (with loko-hauled trains) on the new infrastructure
3. 2010 timetable (with trainsets) on the new infrastructure
4. 2020 timetable (with trainsets) on the new infrastructure

The scenarios are chosen so only one variable is changed at a time. Thus between scenarios 1 and 2 the infrastructure is changed, between 2 and 3 the rolling stock used is changed and between 3 and 4 the timetable used is changed.

Based on the case, the selected indicators are evaluated by comparing indicator results on the case with the result of a microscopic simulation. The microscopic simulation is conducted using RailSys. Input delays (perturbations) for the simulation are generated on the basis of manually and automatically collected track occupation data (Rail Net Denmark RDS data) from the 2010 timetable. Based on this data, a calibrated simulation

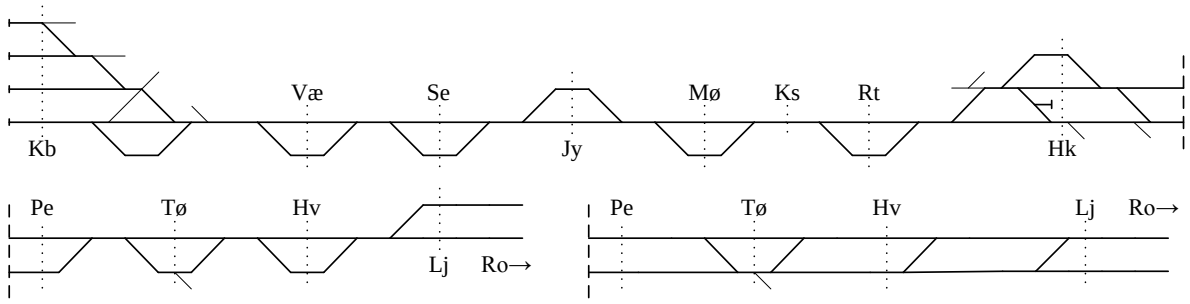


Figure 5.1: Schematic track plan before and after extension. Lower left part: base scenario. Lower right part: future scenario.

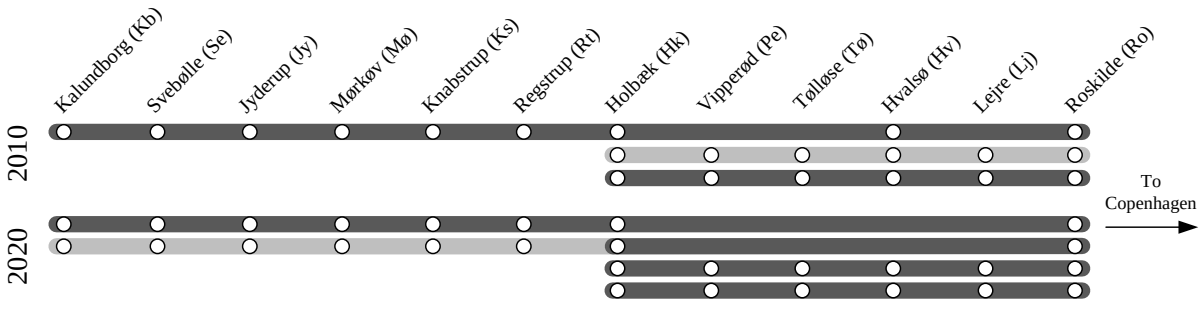


Figure 5.2: Line diagram in 2010 and 2020. Light grey lines: peak hours only.

model can be obtained in accordance with the first calibration procedure described in section 3.3.1. In this context the North West line is considered as a closed system where trains enter with a certain amount of input delay at Roskilde station and exit with a certain amount of exit delay.

In the calibration, input delay distributions are assigned on stations between Roskilde and Kalundborg so the exit delay at Roskilde is equal to the recorded delay. In the simulation of the future infrastructure the calibrated distributions are used to obtain a realistic result. For all simulations the robustness is measured both quantitatively, as punctuality and average delay, and qualitatively by analysing the quantitative results. Especially the stability of the system is important. Following Goverde (2008)'s definition of local stability, the timetable can be considered stable if the exit delay is smaller than the input delay (see section 3.2).

The simulation is used to assess stability and robustness in line with the purpose of the robustness indicators. Resilience of the railway system is not assessed (see section 3.2 for definitions). Therefore only smaller (daily occurring) delays are considered in the simulation. Thus input delays are truncated at 10 minutes and dispatching measures are not used in the simulation.

For the calculation of robustness indicators minimum headway times, buffer times, time supplements and minimum running times are collected from the RailSys model. For the calculation of complexity indices at crossing stations only through-going train routes have been considered.

5.2 Results

In the following two sections, the results of the case study is given. First, in section 5.2.1, robustness indicators results for the case is given. While section 5.2.2 relates the indicator results with the results of the microscopic simulation performed and thus the quality of the indicators.

5.2.1 Indicator results

Figure 5.3 shows some of the indicator results for the four scenarios. From the UIC 406 calculation of the timetables it is seen that the amount of buffer time is significantly lower in scenario 1 compared to scenario 2-4 on the single track section between Lejre and Vipperød, i.e. the infrastructure occupation is lower. This is expected as this is the section that is upgraded to double track. The increase in trains from scenario 1 to 4 is

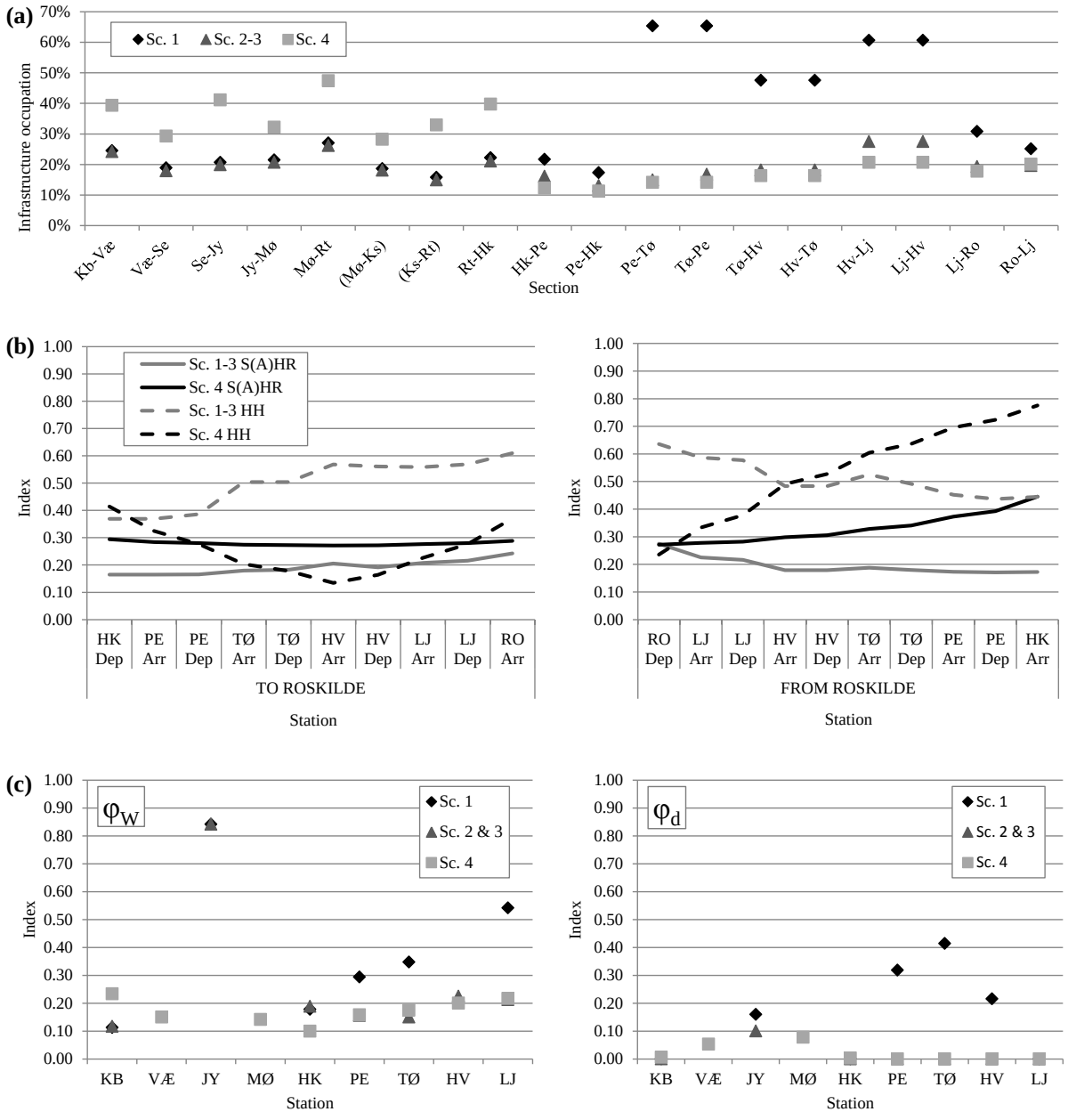


Figure 5.3: (a) UIC 406 infrastructure occupation (b) Heterogeneity indices (c) Complexity indices.

also seen as the buffer times are smaller on the unaltered single track between Holbæk and Kalundborg.

The heterogeneity index, HH, (figure 5.3b) shows that the (relative) heterogeneity in scenario 4 is lower than in scenario 1-3 towards Roskilde and approximately the same in the other direction. The $SAHR$ index shows that the heterogeneity of scenario 4 is higher than in the other scenarios. This is a result of the higher amount of trains operated and thus lower headways times.

In figure 5.3c the complexity indices, φ_W and φ_d, are shown for the four scenarios. Where φ_W is the infrastructure complexity weighted by plan of operation and headways, and φ_d is the timetable complexity taking the infrastructure, timetable, and expected dealys into consideration. Values of φ_n and φ_p (infrastructure complexity) are 1 for Kalundborg, the crossing and transition stations (not depicted). The values of φ_n and φ_p drop to 0.5 in scenarios 2-4 for the crossing stations that is situated on the upgraded line. This is caused by the separation of the two directions. φ_p for Holbæk, where some of the trains terminate, shows that the complexity is increased in scenario 4 due to the increased number of trains. The more detailed complexity indices φ_W and φ_d show that the crossing stations and the transition station Vipperød are of highest concern. This is especially the case with φ_d where these are the only stations where the value is above 0. For Hvalsø-station φ_W indicates no difference among scenarios.

From the analysis of the train patterns, with the use of fix points and risk profiles, the amount of running time supplements is particularly interesting (see figure 5.4). Scenarios 1-3 are operated with high amounts of running time supplements in the excess of 20% on average. This is much higher than Rail Net Denmark's recommended 5% at 120 km/h. Some of the reason for this is scheduled waiting time. However, in scenario 4 the running time supplements are much lower. 6% for the outbound (from Roskilde) Kalundborg trains and only 3% for the inbound. For the Holbæk trains the running time supplement is 8.6% which is in line with Rail Net Denmark's recommendation of 9% at 160 km/h.

The conclusion to be drawn from the robustness indicators are: the crossing stations are of concern in terms of delay propagation and the amount of running time supplements are very low in scenario 4. The single track section between Vipperød and Lejre makes scenario 1 vulnerable in terms of delay propagation. However, there is a large amount of running time supplement. Scenario 4 seems very vulnerable as the amount of supplements is low and the available buffer time between Kalundborg and Holbæk has been decreased.

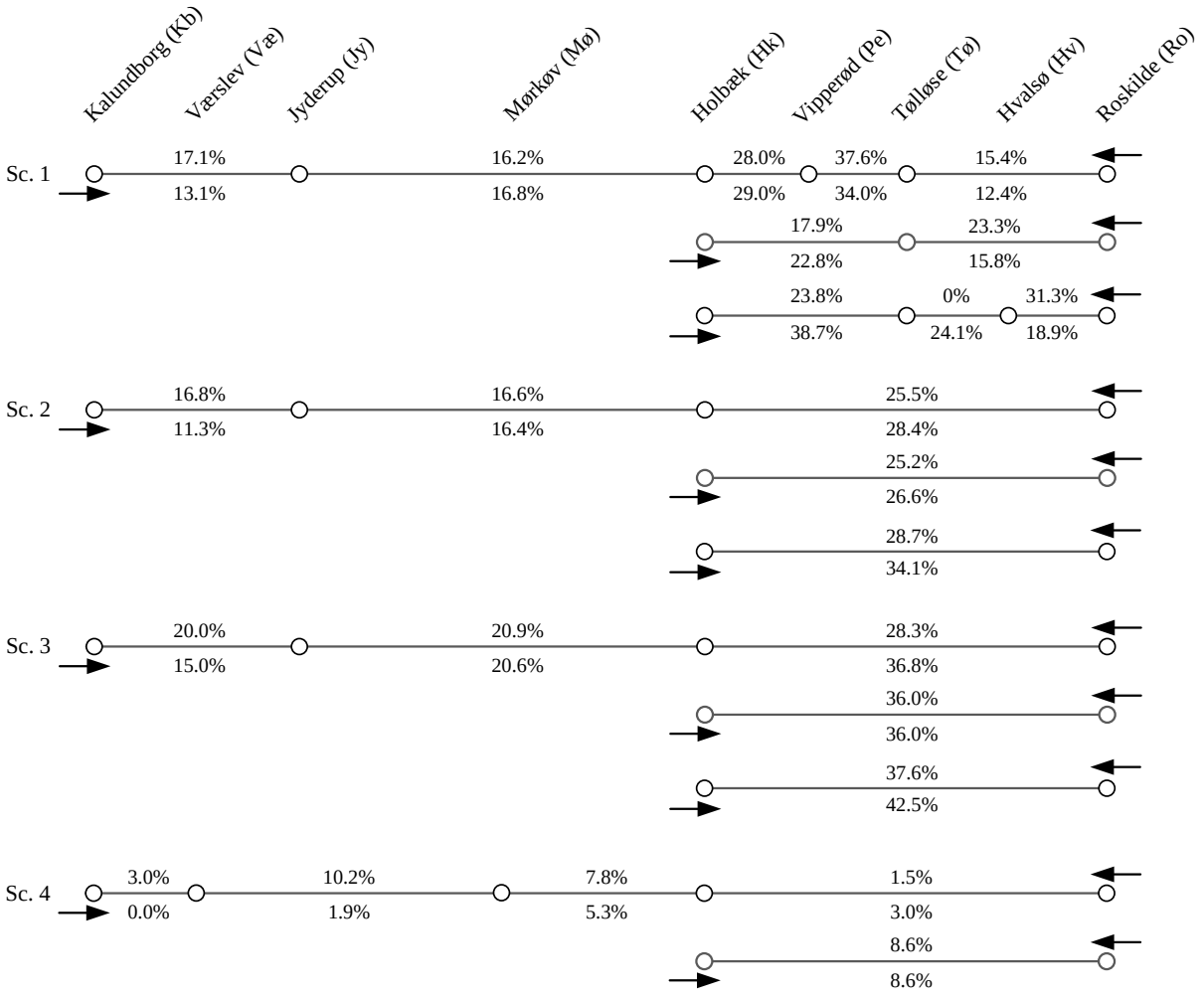


Figure 5.4: Running time supplement between fix points in the four scenarios. Circles denote fix points.

5.2.2 Comparison with simulation results

The delay data for the 2010 timetable show that trains enter the North West line with a fairly high amount of delay, but exit with less and more acceptable delay. Following the definition of stability by Goverde (2008) scenario 1 is stable as the output delay is smaller than the input delay. The simulation of scenario 1 shows that delays increase at the stations, especially the crossing/transition stations between Lejre and Holbæk. This is also identified by φ_d. However, the simulation also shows that the timetable is able to quickly absorb delays due to the high running time supplement. As the infrastructure is upgraded in scenarios 2 and 3, but the timetable is the same, the simulation of these scenarios shows even higher robustness and stability. However, considering the robustness definition by Dewilde (2014) these scenarios are probably not robust as they contain a high amount of unused supplements (inefficient) leading to scheduled delay of passengers (long travel times). Thus the scenarios are stable according to the definition by Goverde (2008), but lack efficiency.

In terms of scenario 4, the combination of low running time supplements and a high amount of input delay at Roskilde station makes the scenario less robust than the others. The timetable is still stable as the output delay is a bit smaller than the input delay. As expected from the indicators the problem is the Kalundborg trains, where the timetable is unstable on the trip from Kalundborg to Holbæk (delay is added). The trains to and from Holbæk on the contrary perform very well, but are also operated with a higher amount of running time supplement as described above.

To conclude, the following indicators compared well to the results of the simulations: infrastructure complexity φ_W, timetable complexity φ_d, UIC 406, and train path risk profiles. A link between results of the simulation and the heterogeneity indices may also exist, but is difficult to identify due to the difference in running time supplements. Furthermore, due to the high amount of running time supplements, delays can be absorbed quickly thus reducing the risk of delay propagation. A more homogeneous distribution of buffer times does therefore not have a significant effect on robustness as delay rarely propagate. Infrastructure complexities φ_n and φ_p captures the change when going from single to double track and the increase in trains at Holbæk (φ_p), but otherwise none. These two indicators are therefore more valuable in infrastructure planning than in timetabling.

5.3 Summary and conclusions

In this chapter, five types of indicators in the form of infrastructure occupation rates (UIC 406) (see section 4.1), heterogeneity indices HH and $SAHR$ (section 4.2), four infrastructure complexity indices (section 4.6.1), a timetable complexity index (section 4.6.2.5), and train path risk profiles (section 4.4.1) have been compared to a microscopic RailSys simulation. These indicators can theoretically measure the robustness of timetables either directly or indirectly based on the distribution and size of initial delays, time supplements, and buffer times as described in chapter 4.

From the comparison conducted in this chapter, it can be concluded that the most detailed infrastructure complexity index, the timetable complexity index, the UIC 406 results and the train path risk profiles performed best in terms of capturing changes in robustness. However, it was not possible to uncover semantics of the indicators due to a fairly small number of scenarios being evaluated. Yet, the results seem to suggest that the timetable complexity, φ_d, for a station should be no more than approx. 0.01 if the operation at the stations is to be considered robust.

To uncover the actual semantics of the indicators better, significantly more infrastructure and timetable scenarios are needed to obtain a larger data set. This would likely make it possible to recommend values of indicators to IMs and RUs for timetable and infrastructure planning purposes. Furthermore, it can be implemented in optimisation models to stop optimisation when a given robustness criterion has been met and other parameters can then subsequently be optimised. In chapter 10 the (future) potential of robustness indicators is discusses in further detail.

Part II

Capacity models for railway networks

6 Railway capacity – definition and assessment

This chapter is partially based on the following article:

L. W. Jensen, A. Landex, O. A. Nielsen, L. G. Kroon, and M. Schmidt. Strategic assessment of capacity consumption in railway networks: Framework and model. *Submitted to: Transportation Research: Part C*, 2016

A proper introduction into the area of capacity research is done in this chapter, before describing the work done to develop methods for capacity assessment. This chapter thus reviews and describes how to define capacity, what parameters that affect capacity, and existing methods to assess capacity.

The present chapter is organised into the following sections. Section 6.1 introduces aspects and definitions of capacity. In section 6.2 different parameters affecting capacity are described more thoroughly. The difference between line and station capacity is shortly discussed in section 6.3. Subsequently, methods for the assessment of capacity are reviewed in sections 6.4 and 6.5. Especially, the UIC 406 method is relevant in relation to the further work and is thus given much attention in section 6.4. Lastly, a summary of the chapter is given in section 6.6.

6.1 Capacity definition

"Capacity as such does not exist. Railway infrastructure capacity depends on the way it is utilised." (UIC, 2004b)

This statement regarding the infrastructure capacity of railways follows from the fact that railway capacity is not straightforward to derive. The actual number of trains that can be handled depends on several operational parameters as well as characteristics of the infrastructure and rolling stock. At first glance this might not be intuitive. For instance

in road transport, capacity is usually defined simply as the number of (standard) vehicles/lane/h for different categories of road. While the same, to some point, can be done for railways, values would be mere guidelines rather than actual design recommendations. This comes down to two factors:

1. Signalling equipment and layout as well as operational rules impose minimum headways unique to each railway segment

2. Train services usually operated on a line have different stop patterns and are operated according to a timetable to serve passenger demand

In road traffic the minimum headway time between vehicles is usually set to 2 seconds corresponding to the maximum brake reaction time of most drivers (Herrstedt, 2007). In railways, the type of rolling stock used and the design of the given signalling system in place, governs the size of the minimum headway distance (which can be be calculated using the blocking time model as described in section 2.3). Therefore, it is straightforward to calculate the capacity of a given road segment, but not for a railway segment.

If the minimum headway time of a given railway segment is known and the operation is homogeneous (only one train type and stopping pattern is used), the maximum theoretical capacity per hour of the given segment is straightforward to measure according to the formula:

$$trains/h = \frac{3600\ seconds/h}{minimum\ headway\ in\ seconds} \tag{6.1}$$

Using this formula the maximum capacity of a given railway segment can be derived easily when there is only one train type and stopping pattern. However, as stated above, passenger demand must be met in the best way possible. To serve passengers travelling longer distances skip-stop services are introduced. These services usually operate on the same tracks as services stopping at all stations. In addition to different passenger service, freight trains may also be operated on the network. As a result trains run with different speeds in many networks. This speed heterogeneity leads to longer headway times as a fast train has to depart later *not* to catch up with a slower train. Thus only a reduced number of trains per hour can be operated compared to a network with homogeneous train operation.

The problem may to some extend be solved by overtaking facilities, however such facilities are only available at some stations or part of the network with more than two tracks. On the contrary, roads provide continuous overtaking possibilities in most cases.

The dependency between capacity and heterogeneity of train types is expressed by UIC (2004b) as:

"the total number of possible paths in a defined time window, considering the actual path mix or known developments respectively and the IM's own assumptions"

As train services are timetabled there naturally is an expectation from the passengers that this timetable is kept. Thus in line with chapter 3 the timetable needs to be robust against delays. This means that buffer times have to be added, thereby reducing the actual capacity available.

As heterogeneity of train speeds as well as expected robustness has an effect on the capacity available, different capacity statements may be defined. Based on the work by Krueger (1999); UIC (2004b); Landex (2008a); Abril et al. (2008), this thesis defines and uses the following definitions of capacity:

Maximum theoretical capacity: The maximum number of trains that can traverse a given part of the network in a predefined time period. The operation must be completely homogeneous with a single ideal train type (providing the lowest block occupations) with no supplements of any kind.

Maximum capacity: The same as the maximum theoretical capacity, however for the maximum capacity, the actual train mix is taken into account and thus the heterogeneity of the operation. Buffer times between trains are not included, and running and dwell time supplements may, in some cases, also be excluded.

Practical capacity: The maximum capacity that can actually be used to obtain a robust and stable operation. That is, buffer times as well as running and dwell time supplements are included. The actual buffer times and supplements added depend on the robustness required and the variations in train processes and external influences. Practical capacity is denoted fundamental capacity by Landex (2008a).

Used capacity: the amount of practical capacity actually used.

Available capacity: The difference between the used capacity and the practical capacity. The available capacity may be divided into usable and lost capacity. The lost capacity is available capacity that cannot be used as the "free" space in the timetable is not sufficient to hold a train path (which is in demand by an operator). It should be noted that capacity is never truly lost as it will always provide additional robustness and stability against delays in the form of buffer times.

Of these definitions, the practical capacity is the most usable as it gives the amount of trains a system can handle with a specific train order, the capability of the infrastructure and rolling stock, as well as the robustness and stability expected.

However, it is also more difficult to derive the practical capacity than the maximum capacity as the necessary (critical) buffer times have to be estimated to obtain the expected level of robustness. This can for instance be done by simulating delays in the given railway system. Alternatively, the maximum capacity can be estimated, and the practical capacity can subsequently be stated relativity to the maximum capacity based on a recommended percentage.

The latter is simpler and easier to carry out compared to a simulation of the system. However, the practical capacity will thus be estimated based on a general recommendation and not the system itself. The relative difference between maximum and practical capacity is given as approx. 67% by Krueger (1999) and 60-85% depending on the type of operation by UIC (2004b, 2013) (see also section 6.4.3).

6.2 Capacity parameters and utilisation

In section 6.1, it was described that the capacity of a railway system depends on the (speed) heterogeneity of train types, the robustness against delays, and the capabilities of the infrastructure and rolling stock. In this section, the parameters affecting capacity will be described more thoroughly as well as how capacity can be utilised differently depending on operational parameters.

Landex (2008a) lists the parameters that determine the capacity of a railway system as the infrastructure, timetable, rolling stock, train processes, and external influences as depicted in figure 6.1.

The infrastructure's effect on the capacity depends on the number of tracks available, (possible) route conflicts in junctions, maximum allowed speed(s), and the signalling system in the form of block divisions and how they are occupied (see section 2.3).

The maximum speed, acceleration, and braking capabilities of the rolling stock used determines the minimum running times of trains which affects the block occupation times. Furthermore, the door layout (size, distribution, and number) and interior layout of trains affect the speed of passenger exchange at stations. Slow exchange speed results in longer dwell times and thus block occupation times at stations.

Variability in train processes and external factors may result in disturbances and disruptions unique to each system as described in section 3.1. To absorb these and obtain a robust and stable system time supplements and buffer times are added to the timetable.

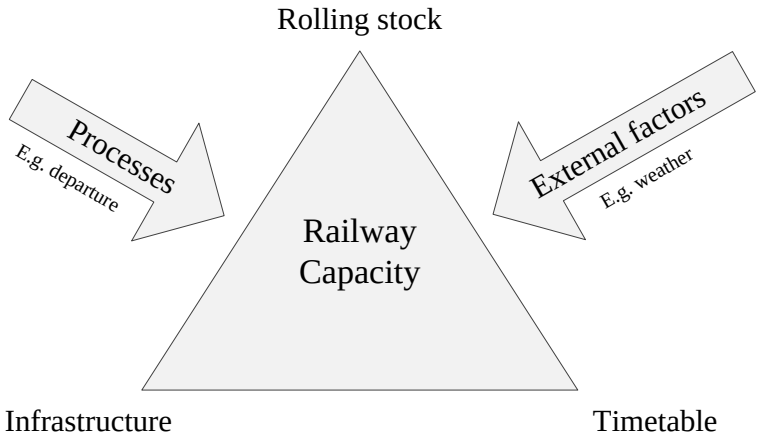


Figure 6.1: Parameters affecting capacity (Landex, 2008a).

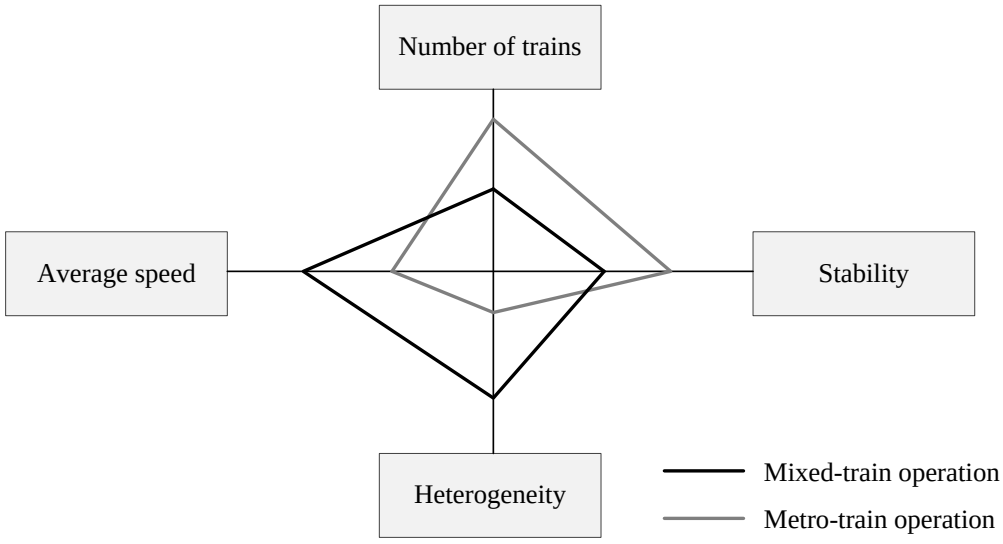


Figure 6.2: The capacity balance (UIC, 2004b). The chords show how the capacity is utilised for two different types of operation.

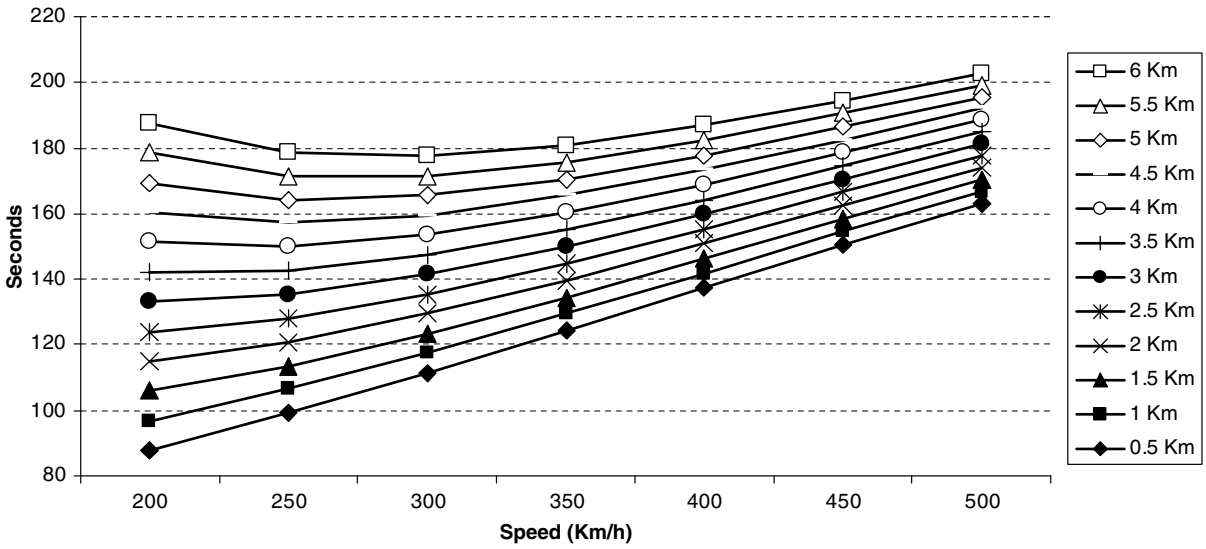


Figure 6.3: The relation between capacity (minimum headway time), block section length, and operating speed (Abril et al., 2008).

In addition to these supplements, the timetable also defines the provided train services and thus the speed heterogeneity of trains operated. As mentioned in section 6.1 time supplements, buffer times, and train heterogeneity affect the number of trains that can be operated. Therefore, the timetable defines how the infrastructure capacity of a given system is utilised.

The UIC (2004b) defines these operational parameters that affect capacity as the number of trains, stability (and robustness), average speed, and heterogeneity. These four parameters together define how capacity is utilised given the infrastructure, rolling stock, and the probability of disturbances and disruptions as fixed. Different timetables may utilise capacity in different ways according to the four operational parameters. In figure 6.2 the utilisation of infrastructure capacity by two different types of operation is shown. In the following each of the four parameters will be described in the context of figure 6.2 and the two types of operation considered in this figure.

The average speed of trains affects the block occupation times as higher speeds reduce the running time through a block section and thus the block occupation time. However, the braking distance increases proportionally to the square of the speed. Therefore, the approach time to a block increases at increasing speeds. This results in block occupation time that in total is longer leading to longer minimum headways. Exceptions from this occur at lower speeds for long block sections where an increase in speed decrease block occupation times. Figure 6.3 shows calculations done by Abril et al. (2008) that illustrates the relation between block length, average speed, and minimum headway time (capacity).

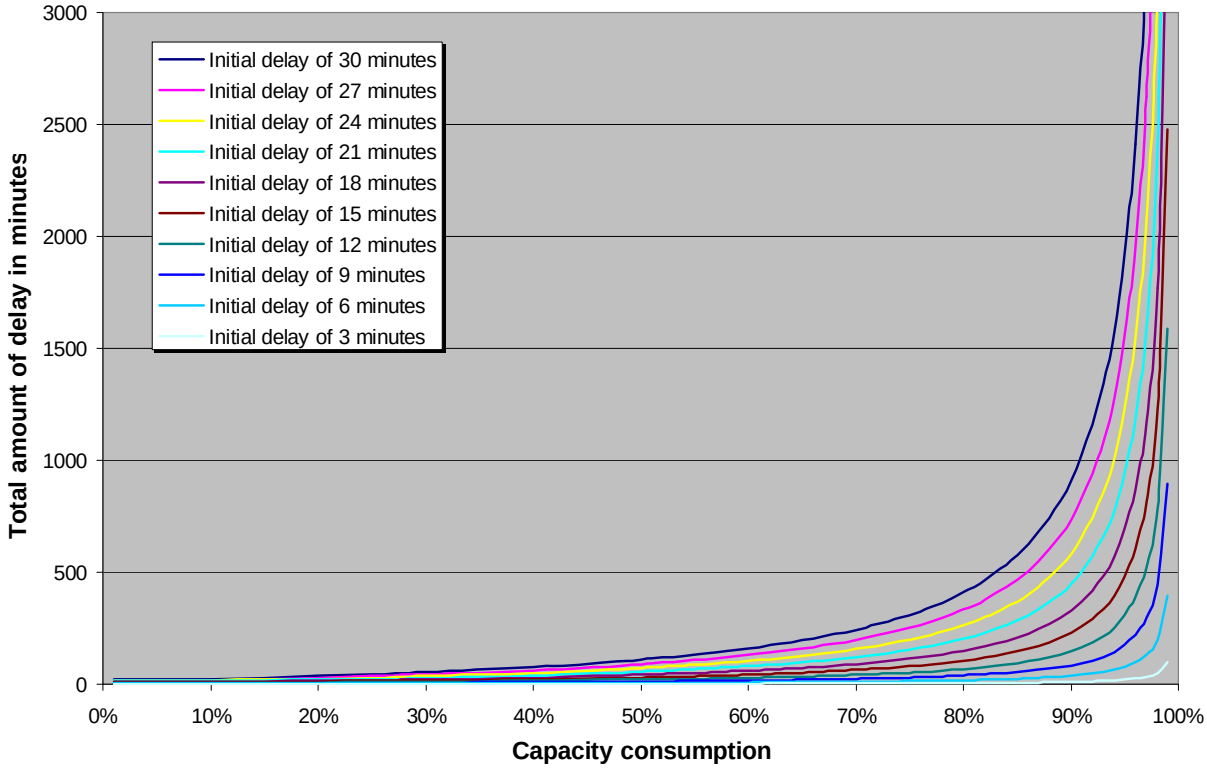


Figure 6.4: The relation between capacity and total delay on a railway line with homogeneous traffic for different amounts of input delay as derived by Landex (2008a).

In figure 6.2 the low average speed in metro-like operation means that minimum headway times are reduced compared to higher speed. For the metro-like operation this is utilised for increased stability and by running more trains. On the contrary, the higher average speed of trains in networks with long distance operation (mixed traffic) results in higher minimum headways and thus a different utilisation of capacity.

As stated earlier a robust and stable system can be obtained by adding running and dwell time supplements and buffer times. By doing so the minimum headway times are increased as time supplements result in increased block occupations times and buffer times has to be inserted between trains. The increased minimum headway times mean that less trains can be operated if heterogeneity and average speed of train services are left unchanged.

As described in section 4.6.2.1 arrival delays may be described by a negative exponential distribution. Thus there is a non-linear relationship between stability/robustness and time supplements and buffer times added at critical locations (where delay is likely to arise). This can furthermore be confirmed by the proof done by Landex (2008a) that shows the non-linear relationship between capacity and total delay for different amounts of input delay as depicted in figure 6.4. Large buffer times and time supplements there-

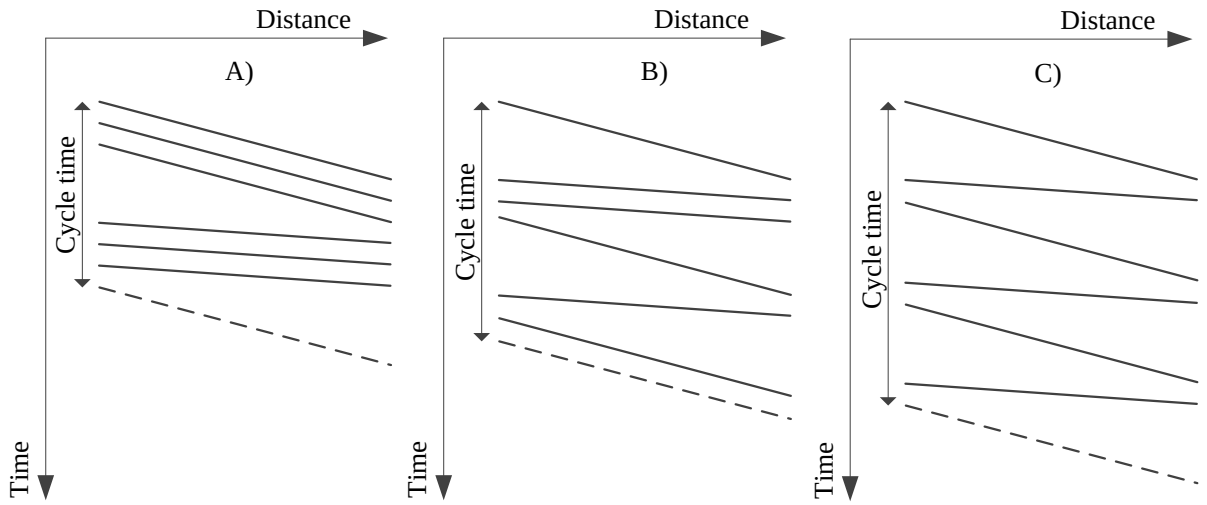


Figure 6.5: Three different orders (sequences) of the same set of six trains divided on three slow and three fast trains. In A) the two train types are bundled (as homogeneously as possible), and thus this sequence consumes less capacity than C) where the sequence is completely heterogeneous.

fore consumes a high amount of capacity with little effect on achieved robustness in daily operation. Buffer times and time supplements should therefore be optimised to obtain the required robustness while saving capacity and shorten travel times.

A high heterogeneity in speed leads to longer line headways between trains of different speeds. This is illustrated in figure 6.5 where (A) shows a cyclic sequence of six trains with trains of the same type bundled. This leads to a low cycle time (the time the train sequence occupies the infrastructure) compared to the most heterogeneous case (C) where no bundling of train types is done. In (C) train types are thus operated alternately. This is very common as it allows different train services to operate in equal intervals. (B) illustrates a scenario with a heterogeneity between (A) and (C). In figure 6.2 the metro-like operation has a low heterogeneity to allow for a higher frequency (number of trains). The mixed traffic operation has a high heterogeneity due to fast and slow trains that serve both short and long distance passengers as well as freight. To obtain this lesser stability and/or frequency has to be accepted.

Lastly, the number of trains (service frequency) is naturally a significant factor for the utilisation of infrastructure capacity. In the context of figure 6.2, running more trains mean that heterogeneity, block occupation time (on the basis of average speed), robustness/stability or a combination hereof has to be reduced. Thus the metro-like operation utilises capacity to obtain a high frequency by reducing average speed and heterogeneity as travel distances in such a system are short.

As described earlier in this section, there is not necessarily a linear relationship between capacity and heterogeneity, average speed, stability/robustness, and the number of trains. For instance the relation between stability/robustness and capacity is non-linear. UIC (2004b) states that the length of the chords in figure 6.2 denotes the capacity, and the position on each axis the value of each of the four operational parameters. However, this is only true if the scale on each of the four axes reflects the true relationship between capacity and the operational parameter of the axis. As stated above this is not linear for all parameters, why the use of chord lengths to reflect capacity in figure 6.2 should be done with caution. To overcome this problem, Landex (2008a) replaces the chord length with a fifth parameter, in a pyramid shape, that represents the capacity consumed given the value of each of the four parameters.

Besides utilising the capacity in different ways to obtain high heterogeneity, stability (robustness), average speed, or frequency, the infrastructure capacity may also be increased. This can be done by upgrading the signalling system and/or optimising the block sections to reduce critical block occupation time and thereby the minimum line headway times. Of course the infrastructure capacity may also be increased by the costly measures of constructing more tracks and upgrading junctions to avoid conflicting routes.

6.3 Station versus line capacity

In section 2.2.1, it was described how the network can be divided into junctions, open line, stops on open line, and stations. Assessing the available capacity for each of these differs greatly in terms of difficulty and complexity.

For open lines the available capacity can simply be derived based on block occupation times derived using the blocking time model described in section 2.3. For stops on the open line, the blocking time model may also be used. However, acceleration and deceleration of trains complicate running time calculations compared to operation at constant speed. For stops it should also be noted that the acceleration, braking, and dwelling of trains lead to longer block occupation times which may result in higher minimum line headways. However, as described by Kaas (1998b) blocks may be shortened before, at, and after the stop to reduce block occupation times.

At stations it is more complicated, especially at larger stations with complex track layouts. The capacity consumed may be derived fairly simple by using the blocking time model at stations. However, this is based on routes used and the order of trains, thus is does not give any indication of the amount of available capacity. The station may simply

be utilised differently by changing the routes used to lower the capacity consumed or fit more trains.

Finding the optimal routing to utilise station capacity in the best way possible is a very difficult problem. Within timetable optimisation a sequential approach is therefore used (Kroon et al., 2008a). First departure and arrival times are derived in one step and a feasible (and possibly optimal) routing plan is subsequently derived (if it exists). Integrating both for larger networks has not been achieved in the literature.

Acceleration, deceleration, and dwell time for passenger exchange consume a high amount of capacity at stations in addition to the risk of conflicting routes. Especially, at large stations this is a problem as the traffic volume in terms of passengers and trains is high.

Therefore station capacity often constitutes the bottleneck(s) of a railway network (Yuan and Hansen, 2007; Landex and Jensen, 2013; Carey and Carville, 2003). This in turn reduces the amount of capacity that can be utilised on open lines as the minimum headway at stations exceeds the minimum line headway (Harris, 2006).

6.4 The UIC 406 method for assessment of capacity

The UIC 406 method is a widely used method to assess capacity developed by the UIC (International Union of Railways) in their leaflet 406 (UIC, 2004b, 2013). The capacity assessment model developed in chapter 7 is an extension of this method. The leaflet describes how to assess the percentual capacity consumed on a piece of infrastructure based on a given timetable, using timetable compression and network decomposition. Based on the capacity consumption estimated it can be derived if there is any residual (available) capacity in (parts of) the network and whether this can be used (usable capacity).

6.4.1 Network division

Based on the network in question and the timetable to be assessed, the network is decomposed into line sections which are to be evaluated separately. UIC (2004b, 2013) recommends that the network is divided into sections where the following does *not* change:

- the number of trains and/or mix of train types fundamentally
- infrastructure and signalling characteristics and conditions

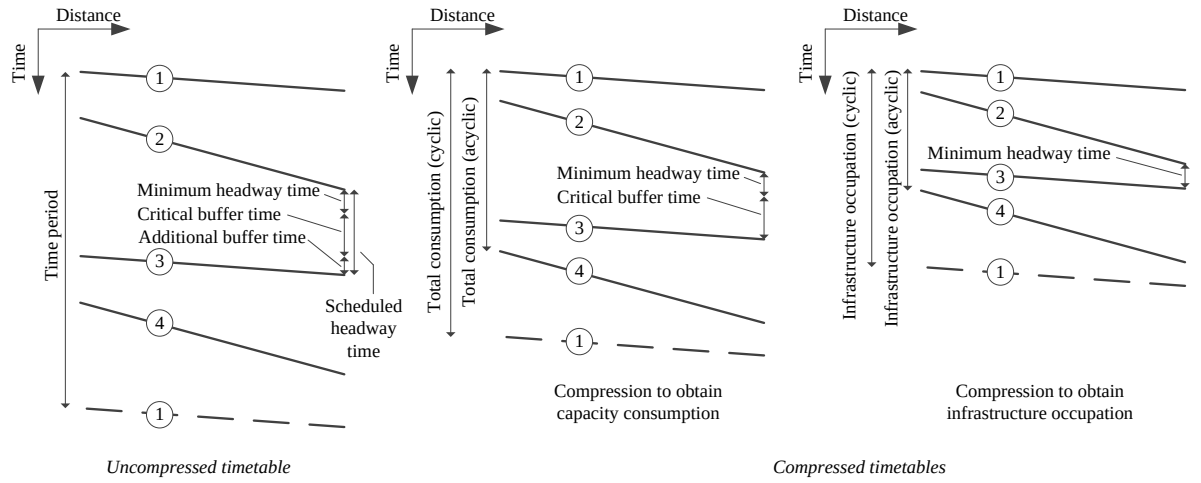


Figure 6.6: Timetable compression to obtain infrastructure occupation and capacity consumption of a train sequence.

Thus the network should be divided at junctions, transition stations, overtaking stations, crossing stations, terminal stations, and stations where the signalling system changes.

However, many stations may provide overtaking or crossing facilities without them being used. Therefore, the network should not be divided here. However, if the timetable is subsequently changed and one of the overtaking stations not used before is now used, the capacity consumed may decrease as the network has to be divided differently. This problem is described more detailed in section 6.4.4 where other deficiencies of the UIC 406 method are also described and discussed.

6.4.2 Timetable compression

For each line section, the UIC method uses timetable compression as depicted in figure 6.6 based on the blocking time model (see section 2.3). Figure 6.6 shows three graphical timetables of a sequence of trains $(1, 2, 3, 4)$. The dashed line defines train 1 in the next cycle (period). If an acyclic timetable is considered, train 1 marked by the dashed line may be neglected.

The leftmost timetable in figure 6.6 is the uncompressed timetable which corresponds to the timetable put into operation. The trains are spaced in time by a (scheduled) headway time consisting of the minimum allowed headway time, the critical buffer time, and additional buffer time.

The **critical buffer time** is the slack time deemed necessary by planners to reduce the risk of delay propagation between trains. The **additional buffer time** is a consequence

of other timetable planning decisions that may include synchronisation at larger stations, heterogeneity in train running times, constraints in other parts of the network, or simply spare capacity.

As the leftover capacity (both usable and unusable) in the timetable is expressed by the additional buffer times. The capacity consumption may be found by compressing the timetable, thus reducing the headway to the minimum allowed plus critical buffer times. This is depicted in figure 6.6 (center) where the (percentual) capacity consumption can be found by the ratio between the total consumption time and the time period considered.

In cyclic operation, the total consumption time is the cycle time of the compressed timetable (see figure 6.6 (center)). While in acyclic operation the total consumption time is the duration from departure of the first train to the departure of the last train.

As figure 6.6 (right) shows, the timetable may be compressed even further by reducing the headway times to the minimum allowed. The time used by the sequence is then defined as the infrastructure occupation, where the ratio between the infrastructure occupation and the considered time period is the infrastructure occupation rate. In cyclic operation the infrastructure occupation is equal to the minimum cycle time.

6.4.3 Recommended values of infrastructure occupation

From a deterministic point of view, the infrastructure occupation rate should be below 100% for the sequence to be feasible as 100% denotes the maximum capacity. However, as delays do occur even values of infrastructure occupation below 100% may be infeasible (due to the lack of buffer times). Thus it is difficult to provide maximum feasible (practical) occupation rates as these differ from network to network, as also mentioned in section 6.1.

UIC (2004b, 2013) recommends that a maximum of 60% to 85% of the maximal capacity should be used. The values are maximum occupation rates recommended for the division into line sections and span over three types of operation in peak hours and off-peak hours (daytime operation) respectively. These six values are listed in table 6.1, where the highest infrastructure occupation rates allowed are for dedicated (sub)urban passenger traffic with 85% in peak hours. This is possible as the high frequencies in such a system makes it possible to cancel trains when larger delays occur with minimal effect on passengers. For mixed-traffic and high-speed operation the recommendation is 75% and 60% in peak hours and off-peak hours, respectively, for both types of operation. However, UIC (2004b) notes that the maximum infrastructure occupation rate for mixed-traffic operation may be increased if only a few number of trains are operated with a

Table 6.1: Maximum values for infrastructure occupation on line sections as recommended by UIC (2004b, 2013).

Line type	Peak hour	Off-peak
Dedicated suburban passenger traffic	85%	70%
Dedicated high-speed line	75%	60%
Mixed-traffic lines	75%	60%

high heterogeneity. This probably stems from the fact that a high heterogeneity results in additional buffer times. These buffer times are inherently spread out between slow and fast trains in the beginning or end of a section thus providing additional robustness. The recommended values in table 6.1 have been found using simulations for a series of cases in, among others, Germany, the Netherlands, and Sweden. UIC (2004a) describes some of this work.

The recommended values for infrastructure occupation rates only provide a guideline due to the broad categorisation, as similar types of operation might differ in the characteristics and behaviour of infrastructure, rolling stock and passengers.

It is, therefore, recommended by this thesis to use capacity consumption values when assessing the feasibility of a sequence of trains (or timetable). I.e. a capacity consumption value below 100% means that the sequence is feasible in practice as 100% denotes the maximum practical capacity (includes buffer times for robustness against delays).

However, to produce capacity consumption values, critical buffer times have to be be estimated. As this is not necessary for obtaining occupation rates, these are therefore easier derived than capacity consumption values.

In case the timetable compression reveals any leftover capacity available, UIC (2004b) suggests the following simple enrichment procedure to determine how many trains that can be added. First, a train in demand may be added if all the line sections in the train's route have available capacity. After the addition of the train, the timetable compression is done again to reveal if there is still available capacity. If so another train may be added. If not, the enrichment procedure is stopped and the critical line section is marked as congested according to the original method (UIC, 2004b), while the updated method (UIC, 2013) suggests to remove the added train again. UIC (2013) suggests that long-distance trains are added/evaluated first, followed by short-distance trains.

6.4.4 Critique of the method

While the UIC 406 method is a widely used method, it has also been criticised for a number of factors some of which have already been mentioned. A general criticism raised

by Lindner (2011) is that the 21 pages of the first version (UIC, 2004b) do not contain enough information to enable the planner to consistently handle all special cases arising in the division of lines and the timetable compression. However, additional details on the process and special cases have been given in the revision of the method (UIC, 2013).

In addition to this, three main problems have been described by Landex (2008a); Lindner (2011):

1. Short line divisions cause unrealistic values of infrastructure occupation
2. The paradox of adding a train leading to a *decrease* in infrastructure occupation
3. Station capacity cannot be handled due to the lack of recommendations for divisions of the station infrastructure, recommended infrastructure occupation rates, and dependencies on the routes used

Landex (2008a) describes how the division into line sections greatly influence results, especially in networks with heterogeneous operation. For instance a line between two bigger stations may be divided into two sections if there is a station that is occasionally used for overtaking. Each of these two sections will each have a maximum infrastructure occupation rate less than the two sections combined. This might give a false indication that there is more available capacity than there actually is as trains cannot necessarily turn around at the station dividing the two line sections. Thus this is not known before the enrichment procedure is started.

This problem may also be illustrated by figure 6.7 which is based on UIC 406 calculations done for the case presented in chapter 5 (based on Jensen et al. (2014)). In figure 6.7, three divisions into line sections are shown; in the top at every crossing or transition station, in the middle only at used crossing/transition stations (UIC 406 recommendation), and in the bottom the whole line. As seen from the figure, the maximum occupation rate for a division is not the same across the three divisions, but rather a greater number when the number of line sections decreases. Thus maximum occupation rate for the top division is 60.6 %, 69.6% for the middle division, and 74.4% for the division depicted in the bottom of figure 6.7.

As mentioned earlier adding new trains to a line may impose that a new division of line sections has to be done. For instance the line has to be divided at an overtaking or crossing station now used. As stated by Landex (2008a), this creates a paradox as adding a train may result in lower occupation values, contrary to the expectation. Landex (2008a,b) recommends processes for partitioning double and single track lines to obtain consistent divisions of the infrastructure and thus consistent results.

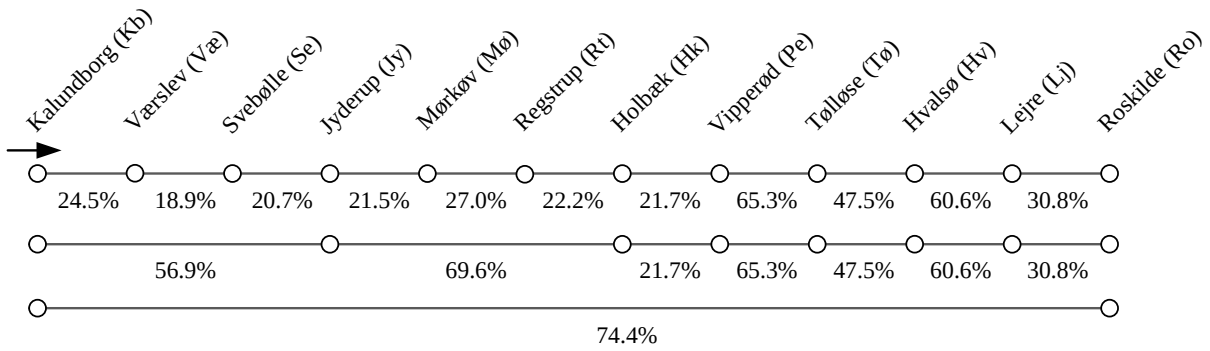


Figure 6.7: Different infrastructure occupation rates depending on division.

At stations, the original UIC 406 method cannot be used, as it does not provide any recommendation as how to divide the infrastructure and the allowed maximum infrastructure occupation rates (Lindner, 2011). Furthermore, the UIC 406 method does not provide the methodology to estimate the available capacity independently of the routes used. This problem was described in section 6.3 regarding utilisation of capacity at station. For instance the UIC 406 method is based on a given route usage. This usage may entail that many other routes cannot be used and may result in all the capacity being used. However, if the route usage is changed, to allow for more concurrent routes, this may decrease capacity usage as described by Lindner (2011). And this is not captured by the UIC 406 method, for which more advanced routing models have to be used as described in section 6.3.

In the revision of the UIC 406 method from 2013 (UIC, 2013), some of the deficiencies of the original method have been dealt with. Now UIC (2013) recommended to look at entire train routes without decomposition when dealing with long-distance services. This accounts for part of the problem with short line sections exaggerating the amount of available capacity. Furthermore, the new version (UIC, 2013) also suggests how to handle stations by dividing the track layout into smaller sections. It should be noted that these improvements of the method do not contain precise recommendations for infrastructure occupation rates at stations and for routes. More work is still needed to obtain these rates.

6.5 Other methods for capacity assessment

In addition to the UIC 406 method there exist several other methods to assess capacity. These methods may use analytical or numerical methods to evaluate the capacity of a given railway system where:

- Analytical methods use closed form expressions to find deterministic and optimal results
- Numerical methods and simulation are used to estimate results to problems where closed-form expressions cannot be derived or are to cumbersome to use

In the following subsections, methods related to assessment of capacity are reviewed. The methods described may either be based on an analytic or numerical approach, or a combination hereof. As defined in chapter 1, this thesis focuses on assessment in the strategic to early tactical planning phase, with a focus on robustness. Therefore this will also be the focus in the review of methods.

6.5.1 CUI method

The CUI method (capacity utilisation index) is an analytical approach used in Great Britain much similar to the UIC 406 method (Sameni, 2012). However, the division of the network is done in smaller sections compared to the UIC 406 method as depicted in figures 6.8 and 6.9. Furthermore, the timetable compression is not based on a detailed blocking time model but rather empirical values for minimum headway times. As is the case with the UIC 406 method, the CUI method cannot handle stations.

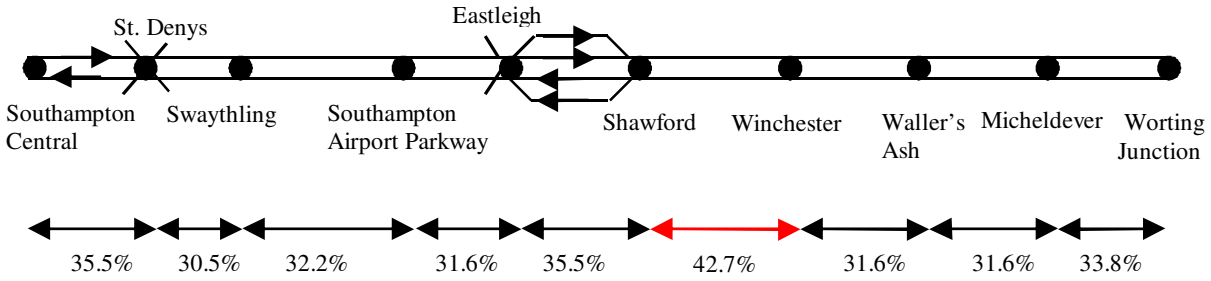


Figure 6.8: CUI analysis for a railway line near Southampton (Sameni, 2012).

From the figures it may be observed that similar sections have different capacity utilisation depending on the method used. This is caused by the fact that link headways are used in the UIC 406 method and headways at nodes (stations) are used for the CUI method. For instance at Shawford station the change between quadruple and double track impose longer headway times by the CUI method and thus a higher capacity utilisation (Sameni, 2012). In practice this means that additional slack (buffer time) is added at the station.

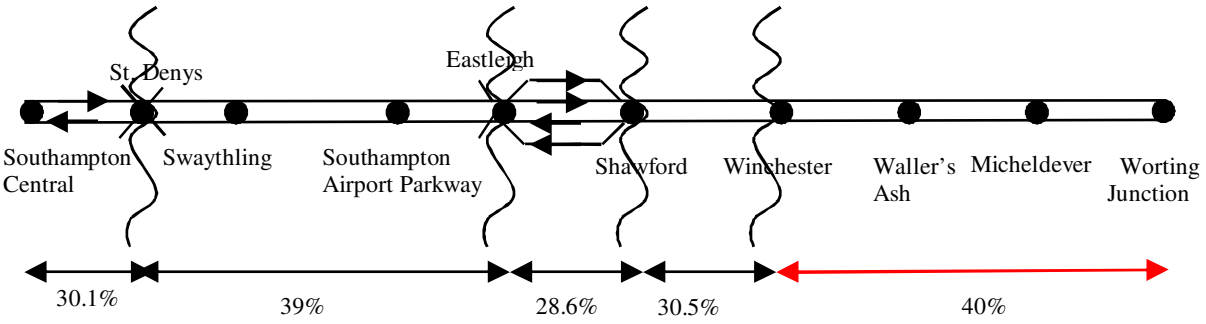


Figure 6.9: UIC 406 analysis for a railway line near Southampton (Sameni, 2012).

Similarly, to the work done by the UIC on stations. There is also ongoing work on how to handle nodes (stations) using the CUI method, including the identification of suitable limits for CUI node values (Armstrong et al., 2013).

6.5.2 Queueing models

Queueing models can be used to assess railway systems without a predefined timetable. For instance Wendler (2007) uses a M/SM/1/∞ queueing model to predict the scheduled waiting time (SWT) on a railway line (defined as the track between two large junctions). This is the scheduled difference between the requested time of arrival to the realised time of arrival. A high SWT indicates congestion and can therefore be used to identify bottlenecks (high capacity consumption). While decomposition is proposed for analysing larger networks; Wendler (2007) only considers a single railway line.

Huisman et al. (2002) analyse a network in the Netherlands using a product form queueing network to assess the performance in the form of mean delay. Input for the model is a number of train classes, each to be operated one or several times per hour. In the model, station tracks are modelled as multi-server queues, while line tracks are modelled as single-server queues. However, in the approach by Huisman et al. (2002) a number of assumptions and approximations have been made to simplify the queueing model. One is that each train is allowed to use all platform tracks. This will rarely be the case, but implementing this in a queueing model is rather complicated as stated by Huisman et al. (2002). This inflexibility displays one of the disadvantages of queueing models when applied to the complicated nature of railway systems compared to i.e. simulation models.

6.5.3 Optimisation

Optimisation models are yet another method that can be used to assess capacity of a railway system. This can either be done directly to assess the number of trains a network

can handle, or indirectly by using the Train Timetabling Problem (TTP) to estimate if there is capacity for a given plan of operation (number and type of trains).

6.5.3.1 Direct

The work of Burdett and Kozan (2006); de Kort et al. (2003); Mussone and Wolfler Calvo (2013), deals explicitly with the determination of railway capacity in terms of number of trains a network can handle. These approaches are suitable for strategic planning as they require no timetable. de Kort et al. (2003) uses a probabilistic approach using max-plus algebra to determine the capacity of railway infrastructure. The approach is based on the bottleneck approach which determines the critical section (bottleneck) that limits the capacity of the system considered. The approach does not explicitly account for different train types (heterogeneity). However, it is possible to implement this by using the probability that a train will be of a given type.

Inspired by the work of de Kort et al. (2003), Mussone and Wolfler Calvo (2013) presents an optimisation framework to maximize the total number of trains in the railway system based on train conflict probabilities, where different train types can be considered. de Kort et al. (2003); Mussone and Wolfler Calvo (2013) are able to include knock-on delays in the railway system in a simple way.

Common for the approaches by de Kort et al. (2003); Mussone and Wolfler Calvo (2013) is that they do not depend on the exact train sequence, but rather consider a weighted average. Thus, in a network with heterogeneous operation the average absolute capacity is obtained, and this capacity might therefore be increased by bundling trains.

Lastly, Burdett and Kozan (2006) describe and discuss terms for absolute capacity and utilisation levels based on sectional running times of train types, dwell times and proportional mix of train types. Based on these, they propose an optimisation model to maximize the number of trains in a railway network. The model proposed is non-linear and is therefore only solved to a local maximum (Burdett and Kozan, 2006). The approach by Burdett and Kozan (2006) is quite flexible and is able to capture a large amount of operational characteristics. Furthermore, the approach is able to give a lower and an upper bound for capacity.

6.5.3.2 TTP

Contrary to the approaches described in section 6.5.3.1, the main part of optimisation models within railways is concerned with the Train Timetabling Problem (TTP). That is, the problem of creating a feasible (and possibly optimal) timetable given a specific

network and lines/passenger demands. This can indirectly be used to assess capacity based on the number of trains that can be successfully scheduled by such a model.

In the context of the robustness point of view used in this thesis, the work of Dewilde et al. (2014); Kroon et al. (2008b); Liebchen et al. (2010) are examples where efforts are made to take robustness of operation into account. However, these models are mainly suited for the later tactical planning level as opposed to the strategic planning level that this thesis focuses on. Thus they required more input data, e.g. guaranteed transfers.

6.5.4 Simulation of operation

Simulation models are a well-known method to assess a given line or network. Commercial railway operation simulation software is the most well-known of simulation models and includes, for example, RailSys (Bendfeldt et al., 2000) and OpenTrack (Nash and Huerlimann, 2004). Based on a timetable, this type of software can be used to assess the feasibility of a timetable taking robustness into account by simulating delay propagation from input delays. I.e. does the railway system have the capacity to operate the specified timetable? For strategic purposes this type of simulation model is not desirable as it requires a detailed timetable and a microscopic infrastructure model. However, it should be preferred in later planning stages because of its high precision.

6.5.5 Others

One sampling model suitable for strategic planning, however no longer in use, is the SCAN model (Strategic Capacity Analysis for Network) of Kaas (1998a). This is a stochastic (Monte Carlo) simulation model that schedules trains in random order using a synchronous simulation approach in each iteration. As the trains are scheduled randomly no timetable is needed, only the number and type of trains to be scheduled including running times (the plan of operation) as well as the infrastructure. The output of the model is the distribution of scheduled waiting time (SWT) added in each iteration. High amounts of SWT imply that capacity is fully utilised. E.g. SWT is added to fast services to reduce the heterogeneity of the timetable by reducing the operational speed of the fast trains. Likewise slow trains may have been scheduled for more overtakings (adding SWT to their running time) for the timetable to be feasible. It is possible to use the SCAN model in large networks, however robustness (delay) is not considered.

More recently, Lindfeldt (2015) has developed a model named TigerSim that can calculate the delay and SWT for a given traffic scenario (plan of operation) on a double track line. The model can be used to estimate capacity by altering the plan of operation

until the quality of service (SWT and delays) deteriorates. The model generates a set of timetables based on the plan of operation and subsequently simulates these using a macroscopic infrastructure model to produce the output (Lindfeldt, 2015). This model is thus suitable for the strategic planning phase as it does not require a predefined timetable. Furthermore, robustness is considered as the model simulates delays. However, only lines are considered by the model.

In the Netherlands, network capacity consumption and the stability of operations is calculated using the analytic model PETER (Performance Evaluation of Timed Events in Railways) of Goverde (2007). This model is based on max-plus algebra and can obtain solutions fast, even for large networks. However, as is the case with the timetable optimisation methods, this model requires a predefined timetable as input and is, therefore, more suitable at the tactical level.

In section 6.4.3 a span of values for maximum infrastructure occupation rates was listed as given by (UIC, 2004b, 2013). However, as described in section 6.4.3 this only provides a very coarse guideline for a given network. To negate this effect, Goverde et al. (2013) extend the compression method to obtain the dynamic infrastructure occupation under disturbances, contrary to the scheduled infrastructure occupation. Thus measures against delays are taking into account yielding more accurate results.

As described in section 6.2, the capacity depends on the order of trains and thus on the timetable in networks with heterogeneous train types. To deal with this issue of heterogeneous train types Lai et al. (2015) use a standardized unit to denote base train equivalents (BTE) and base train units (BTU) to define capacity on lines in homogeneous way.

Lastly, the parametric model of Krueger (1999) is a model to estimate the capacity of railway lines based on a set of operational and infrastructure parameters. Specifically, the model predicts a curve of train delay based on the number of trains per day according to the formula (Krueger, 1999)

$$\text{Train delay} = A_0 \cdot e^{B \cdot V} \tag{6.2}$$

Where A_0 is a coefficient expressing parameters of infrastructure and operational characteristics of the line, B is a constant, and V is the traffic volume in terms of number of trains per day. This relationship including values of A_0 have been found by performing a series of event based simulations with differentiated infrastructure and operational parameters. The fitted parametric model shows a maximum of 10% deviation from simulation for a line in Canada (Krueger, 1999). The model of Krueger (1999) is valuable

in a strategic planning context. However, only lines can be handled and thus the model cannot assess network capacity without the use of decomposition.

Finally, complexity indices, as described in detail in section 4.6, may be used to compare the capacity of different infrastructure layouts at stations. As mentioned in section 6.3, stations, and the conflicts in the switch zones of stations, often constitute the capacity bottlenecks of a railway network. Thus the ability to evaluate conflicts in station switch zones, using complexity indices, is an valuable tool for the assessment of overall network capacity.

6.6 Summary

In this chapter an introduction to railway capacity has been given. As described in detail in the chapter, railway capacity is not easily defined as it depends on several infrastructure, rolling stock, and operation parameters. These parameters consist of the infrastructure in the form of the number of tracks, line speed allowed, potential conflicts at stations, and the signalling system used. Furthermore, the dynamics (acceleration, deceleration, maximum speed) of rolling stock and the speed heterogeneity of train services, operating speed, and required robustness/stability of the train operation has a large impact on the number of trains that a railway system can handle.

To assess railway capacity several methods ranging from analytical to simulation methods can be used, or a combination hereof. One method widely used, and the basis for the framework and model developed by this thesis (described in section 7), is the UIC 406 method. This method describes how the capacity in a railway network can be assessed by dividing the network into line sections for which a compression of the timetable is done to obtain the capacity consumed for each section in the timetable. While this method is simple, it requires a timetable which is not desirable from a strategic planning perspective. Furthermore, the division of the network may lead to optimistic estimates of available capacity.

7 Framework and model for assessment of capacity consumption in networks

This chapter is based on the following articles:

L. W. Jensen, A. Landex, O. A. Nielsen, L. G. Kroon, and M. Schmidt. Strategic assessment of capacity consumption in railway networks: Framework and model. *Submitted to: Transportation Research: Part C*, 2016

L. W. Jensen, A. Landex, and O. A. Nielsen. Assessment of Stochastic Capacity Consumption in Railway Networks. In *Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis*, Tokyo, 2015

With an ever increasing demand for transport, knowledge of railway capacity is crucial for the railway industry to efficiently utilise and manage the available resources in the best way possible (Abril et al., 2008). However, as chapter 6 describes in detail, the capacity of a railway system cannot be determined in a straightforward way. In particularly, the capacity of a railway system usually has to be assessed based on a timetable. In addition to the considerable time it takes to plan a timetable, an assessment constrained to a particular timetable is not desirable from a strategic planning perspective as the timetable derived will be uncertain due to the long planning horizon.

In this chapter, this thesis therefore develops a framework that can be used to determine the capacity consumed by a set of trains traversing a specific network, independent of a predefined timetable, while taking measures into account to obtain a robust capacity utilisation.

Specifically, the framework extends and improves the UIC 406 method (UIC, 2004b, 2013) to be used in strategic capacity planning with the following properties:

- No predefined timetable is needed (only the service intentions in a plan of operation)
- Networks can be handled, not just lines or line sections

- Estimation of critical buffer times for absorbing delays and thus achieving a robust capacity utilisation

Several methods exist to estimate the capacity consumed by a set of trains. However, none of these methods have all of the properties listed above (as reviewed in sections 6.4 and 6.5). Thus this thesis contributes to the area of capacity research by developing a framework that has all the three properties. In addition to this, an implementation of the framework is developed.

The framework is especially useful for assessing different infrastructure scenarios as the capability of the infrastructure to support different timetables can be analysed easily. Furthermore, the ability to handle networks makes it possible to assess the impact on network capacity from different upgrades, and thus which upgrades has the highest capacity benefit.

The chapter is organised as follows. In section 7.1, related approaches to the assessment of capacity are shortly summarised based on the literature review conducted in chapter 6. In section 7.2, the developed framework is described in detail. Section 7.3 describes the implementation of the developed framework. While in section 7.4, the framework developed by this thesis and the implementation of it is applied on a network in Southwestern Denmark for four different infrastructure scenarios. Main findings are summarised in section 7.5.

7.1 Related approaches

Several methods related to the area of capacity research exist, spanning over various analytical methods, optimisation models, and simulation as described in sections 6.4 and 6.5. Such methods related to strategic capacity assessment are summarised in table 7.1.

The methods listed in table 7.1, have been reviewed with a focus on the ability of the methods to deal with networks, independent of a predefined timetable, while including estimation of measures to obtain robustness against delays. The review shows that existing methods does not possess all three of these abilities as also summarised in table 7.1. This thesis has therefore developed a framework that has all three of these abilities as further described in section 7.2. For an in-depth description of the UIC 406 method see section 6.4 and section 6.5 for a description of the other methods listed in table 7.1.

Table 7.1: Summary of approaches related to capacity assessment and their ability to do so (denoted by •) in networks, independent of the timetable and by including the aspect of robust capacity utilisation. Parenthesis imply that only some methods within each group of methods have the ability specified, or that the ability specified is only partially obtained.

Method (type)	Timetable independent	Network assessment	Robustness
UIC 406			(•)
Timetable optimisation		(•)	(•)
Capacity optimisation	(•)	•	(•)
Queuing models	•		(•)
Simulation		•	•
Others	(•)	(•)	(•)

7.2 Framework for strategic assessment of capacity consumption

In this section, the developed framework is described in detail. A model that implements the framework is given in section 7.3.

As mentioned in the chapter introduction, this framework is an extension of the UIC 406 method (UIC, 2004b, 2013) which uses timetable compression (see section 6.4.2).

The UIC 406 method uses decomposition into line sections as described in section 6.4.1. However, a division of the network may lead to ambiguous results when done inconsistently or in special cases (see section 6.4.4). This thesis therefore recommends to consider the whole network (or as a minimum the whole route of a train) and thereby no division of the network should be made.

When considering the whole network, there will, however, be a large difference in maximum practical infrastructure occupation rates for different networks as their characteristics may differ significantly. Thus no precise recommendations can, presently, be made for maximum infrastructure occupation rates when compressed timetables for whole networks are considered.

This thesis therefore recommends using capacity consumption values when assessing the feasibility of the sequences considered in a network. I.e. a capacity consumption value below 100% means that the sequence is feasible in practice (i.e. it includes buffer times for robustness against delays).

To produce capacity consumption values, critical buffer times have to be estimated as described in section 6.4.2. As this is not necessary for obtaining occupation rates, these

are therefore easier derived than capacity consumption values. This is also a reason for using the network divisions in the UIC 406 method.

As the infrastructure occupation and capacity consumption differs according to the train sequence (the order of trains) in railway networks with heterogeneous train types and/or crossing routes, this thesis proposes to assess all or a representative sample of train sequences. Specifically, the framework procedures described in sections 7.2.1 and 7.2.2 are therefore proposed for the assessment of infrastructure occupation and capacity consumption for strategic purposes in networks (summarised in figure 7.1).

7.2.1 Deterministic – infrastructure occupation

It is now described how the infrastructure occupation is derived in the deterministic case, that is delays are not taken into account. Subsequently, section 7.2.2 focus on the stochastic case, where delays are considered.

From the set of trains, train sequences are generated and the infrastructure occupation for each of these sequences is derived. In case that the number of trains (and thus the number of possible sequences is small), all possible sequences can be generated. In other cases, a random sample of sequences can be generated. In both cases, this thesis proposes that undesirable sequences are excluded, that is sequences which do not make sense from a train operations perspective and would thus never be implemented. E.g. bundling of the same train services results in long waiting time for passengers, which is not desirable, and should therefore only be done in bottlenecks of the network.

As described above this thesis considers not only one specific sequence, but different permutations of train sequences. Therefore, a predefined timetable is not needed as the heterogeneity of train types is handled by examining multiple sequences.

As multiple sequences are considered, multiple values of infrastructure occupation will be generated (when train types are heterogeneous). These values may be sorted in order of infrastructure occupation rate to produce a cumulative distribution of infrastructure occupation as shown in the lower part of figure 7.1.

In the example in figure 7.1, the y-axis denotes the cumulative percentage. This percentage shows how many of the desirable train sequences fit within the infrastructure occupation depicted on the x-axis. That is if all sequences assessed are sorted by infrastructure occupation, then the y-value of a point on the graph denotes the percentage of all assessed sequences that have an infrastructure occupation equal to the x-value or lower. As described in section 6.4.3, train sequences yielding occupation values of 100% or close will not be feasible in practice due to lack of buffer times.

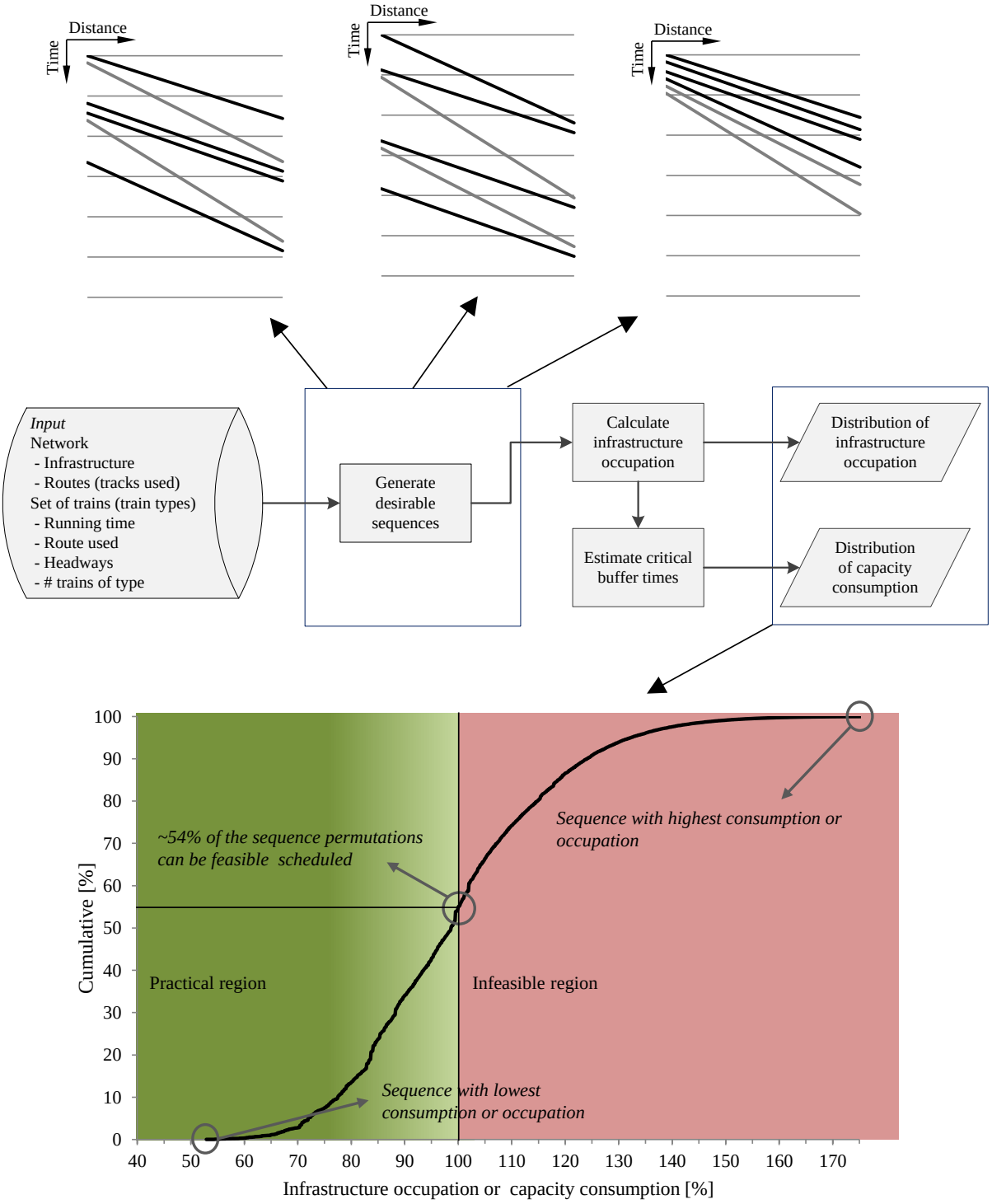


Figure 7.1: Developed framework for strategic capacity assessment.

7.2.2 Stochastic extension – capacity consumption

The previous section described the outline for obtaining a distribution of infrastructure occupation based on permutations of train sequences. To obtain the distribution of capacity consumption, the critical buffer times have to be estimated as described in section 6.4.3.

This can be done by estimating the delay propagation for each compressed timetable based on each of the generated sequences and a sample of initial delays. The delay propagation estimated yields the critical positions in the network and the compressed timetable, where critical buffer times have to be put in. Thus the extended time used by a sequence due to delay propagation is the total consumption time.

Based on the total consumption time of each sequence, a distribution of capacity consumption is produced as multiple values are produced due to the heterogeneity of the different train sequences.

When the graph in figure 7.1 is observed as capacity consumption, then train sequences yielding capacity consumption values below 100% are feasible and sequences above are infeasible.

In figure 7.1, 54% of the sequences can be scheduled feasibly. Thus if a given set of trains is randomly sent into the network considered, there is a probability of 54% that all trains have departed within the time period considered.

7.2.3 Number of sequences

When using this framework it should be noted that the number of possible unique permutations of sequences grows in factorial order of the number of trains. For cyclic and acyclic timetables, there are $(n-1)!$ and $n!$ possible permutations of n train runs, respectively. For instance with 18 trains in acyclic operation the number of permutations is more than six quadrillions (6,402,373,705,728,000). However, in many cases some of the trains are of the same type, and a multiset of trains are thus considered rather than a distinct set of trains. The number of permutations of a multiset can be found using multinomial coefficients (Brualdi, 1977):

$$\binom{n}{m_1, m_2, ..., m_l} = \frac{n!}{m_1! m_2! ... m_l!} \tag{7.1}$$

Where $m_1, m_2, ..., m_l$ are the multiplicities of each train type, l is the number of train types, and n is the total number of trains ($n = \sum m_1, m_2, ..., m_l$).

However, the majority of these sequence permutations is not desirable from a demand point of view and should therefore be discarded as mentioned earlier.

7.3 Implementation of framework

The framework described in section 7.2 defines the overall procedures to obtain capacity estimates independent of a predefined timetable for networks while accounting for robustness. In this section, the model implementation of the framework, developed by this thesis, is described. In section 7.3.1 the infrastructure model used is described and subsequently each of the following steps of the model (sections 7.3.2 through 7.3.4):

1. Generation of train sequences, and exclusion of undesirable sequences
2. Calculation of infrastructure occupation for each sequence generated and generation of a compressed timetable
3. Calculation of capacity consumption using simulation of each compressed timetable

See also appendix B.1 where the model is summarised visually in a flowchart.

7.3.1 Infrastructure model and operational input

In the model developed by this thesis, the infrastructure is represented and stored using a two-layer approach where the first layer ensures precision and the second layer fast computation times.

In the first layer train running and minimum headway times are calculated based on route and train characteristics with as high precision as possible. Preferably a microscopic infrastructure model should be used to ensure feasibility at the track level. This is essential as the output quality of any given model will never be better than its input. However, in some cases it might not be possible to use a microscopic infrastructure model for the whole network as the required input is not available.

E.g. parts of a given network might not exist yet. In this case the existing parts of the network can be modelled using a microscopic infrastructure model while future extensions can be modelled using a simpler model or by simply using the desired headway times and speed profile(s).

In the second layer the data derived in the first layer is aggregated and stored in a mesoscopic graph data structure. While containing the most important information, the

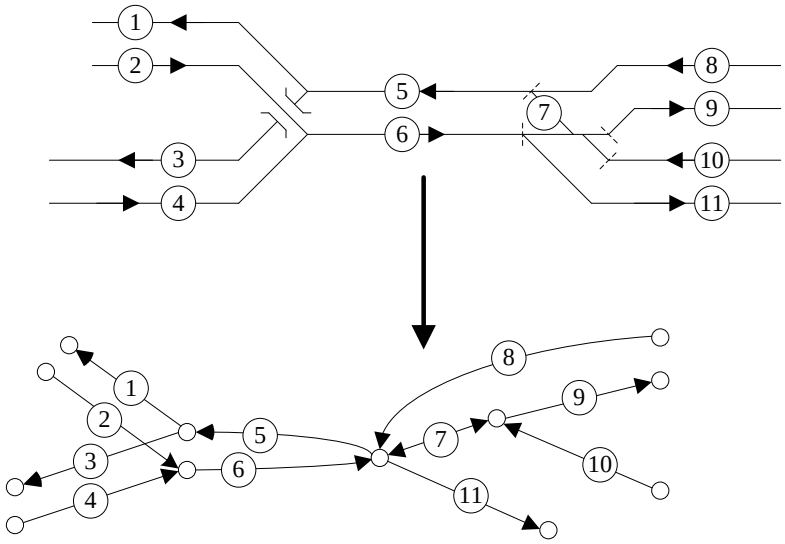


Figure 7.2: Transformation of a simple arbitrary network into the mesoscopic infrastructure model. Schematic drawing at the top, model representation at the bottom. Dashed lines at top show the boundaries of the junction.

size (vertices and edges) of the mesoscopic graph is reduced compared to a microscopic graph yielding faster computation and the possibility to handle larger networks.

This two-layer approach used is related to the micro–macro approach studied and used by other authors as described in section 2.2.2. E.g. Gille et al. (2010); Schlechte et al. (2011) use this approach to exploit the pros and cons of macroscopic, mesocopic and microscopic infrastructure models. Furthermore, Gille et al. (2010); Schlechte et al. (2011) also study how to transform an infrastructure model from one level of detail to another without losing feasibility.

For the mesoscopic representation of the railway network a graph $G = (V, E)$ is used. In G an edge $e \in E$ can be directed or pseudo bidirectional. A pseudo bidirectional edge can be traversed in both directions and is used to represent at-grade junctions. An at-grade junction is a junction where tracks crossing from opposing directions have the same elevation (level), thus crossing conflicts arise as no bridges or tunnels are put into place to prevent these. In figure 7.2, a transformation of a railway network into the mesoscopic graph is shown.

In the figure, the simplification of a junction as a single edge is shown (denoted 7). In this simplification, the two switches marked by dashed lines to the left of 7 (top) are aggregated as the vertex left of 7 (bottom), likewise the two other locations marked by dashed lines is aggregated to the vertex right of 7. Conflicts only occur on edges, vertices have unlimited capacity. Thus the aggregation does not alter the capacity available.

As an operational input to the model, a set of trains S that will traverse the network is given. Each train $s \in S$ uses a route $r \in R$ and leaves the network at the end of the route. A route consists of a set of continuously connected edges, $E_r \subseteq E$. Each train $s \in S$, have edge running times, T_s. Minimum headway times for each edge $e \in E$ are represented by a $n \times n$ matrix, H_e where n is the number of trains in the network.

7.3.2 Generation of train sequences

As described in section 7.2 desirable train sequences need to be generated, where a train sequence is a set of trains ordered to be served (see figure 6.6 for an example). Two ways of approaching the problem of sequence generation can be used. Either generate all possible sequences of the set of trains or only a random sample. As described in section 7.2, the number of permutations of a set grows in a factorial order. Thus, even if the number of trains in the network is low, the number of possible sequence permutations can be very high and sampling is necessary.

7.3.2.1 All permutations

There exist numerous ways of generating all permutations of a set (Sedgewick, 1977). In the developed implementation, all permutations of a set or multiset are generated by an algorithm that generates all sequences in lexicographical order similar to algorithm L given by Knuth (2005).

7.3.2.2 Random permutations

When the number of trains is too high, usually around 13 ($13! = 6,227,020,800$) or higher, this thesis uses Monte Carlo sampling to obtain random sequences. For generating random sequences, the well-known Fisher-Yates (Knuths) shuffle (Knuth, 1997) based on random numbers is used. This algorithm is efficient as it generates random sequences in linear time in order of the number of trains with constant space complexity (Knuth, 1997).

The necessary sample size of train sequences to obtain a result as close as possible to the real result (all permutations) is determined dynamically during the sampling. This is done by using the mean and standard deviation of the calculated infrastructure occupation values, X, as a convergence criteria. Formally, the following statement is tested to be true during the Monte Carlo sampling of sequences:

$$|sd[X_{(N-N_s)}] - sd[X_N]| \leq \epsilon_{sd} \quad \wedge \quad |E[X_{(N-N_s)}] - E[X_N]| \leq \epsilon_E \tag{7.2}$$

Where X_N is the current sample considered and $X_{(N-N_s)}$ is the previous sample measured against, N_s is the number of sequences drawn between these two samples, $sd[X]$ is the sample standard deviation, $E[X]$ is the sample mean (expected value), and ϵ is the tolerance.

When the statement in eq. 7.2 has been true for q consecutive times during the Monte Carlo sampling, convergence in infrastructure occupation values is assumed. At this point sampling of new train sequences is stopped as the result (mean and standard deviation) has stabilised due to a sufficiently large sample size (law of large numbers).

7.3.2.3 Exclusion of sequences

As mentioned in section 7.2, there is no practical interest in generating train sequences that are not desirable from a demand perspective. In this context, two types of sequence constraints are introduced. One is used to exclude generated sequences with certain combinations, the other to ensure that only sequences with certain combinations are included.

Formally, the two constraint types are defined as:

$$s_a \prec \gamma_{e,exc} \prec s_b \tag{7.3}$$

$$s_a \prec \gamma_{e,inc} \prec s_b \tag{7.4}$$

Where $\gamma_{e,exc}$ is an arbitrary non-empty permutation of the set or subset of (thus it can be of any length greater than 0):

$$\{s \in S \setminus \{s_a, s_b\} \quad | \quad e_{ab} \in R_s\} \tag{7.5}$$

Similarly, $\gamma_{e,inc}$ is an arbitrary permutation of the set or subset of (that can be of any length):

$$\{s \in S \setminus \{s_a, s_b\} \quad | \quad e_{ab} \notin R_s\} \tag{7.6}$$

In eq. 7.3 it is stated that train s_a must precede the sequence $\gamma_{e,exc}$ which in turn must precede train s_b. The sequence $\gamma_{e,exc}$ is any sequence of any length greater than 0 containing trains where edge, e_{ab}, is part of the route, R_s, of each train, s. Edge e_{ab} is also used by s_a and s_b. This constraint type makes it possible to avoid sequences where trains s_a and s_b immediately follow each other on a specific edge in the network.

Similarly, the constraint in eq. 7.4 can be used to ensure that train s_b follows s_a at edge e_{ab}. $\gamma_{e,inc}$ is any sequence of trains not using edge, e_{ab}, (not part of R_s) and can be

of any length. This sequence constraint makes it possible to model coupling of trains at a specific edge (station) in the network by ensuring that conflicting trains (trains not to be coupled) are not part of the sequence between s_a and s_b.

In theory also overtakings and turn-arounds can be modelled using this type of constraint. However tests conducted by this thesis show that this does not work well in practice due to a large reduction of the solution space and unrealistic overtakings being generated (bad overtakings resulting in an increase of infrastructure occupation rather than a decrease).

The sequence generated is evaluated and accepted or rejected using the constraints defined to only produce the desirable sequences. It naturally follows that a given set of constraints might produce a very small solution space.

This is not a problem when all permutations are produced. However, in most practical cases the sampling procedure described in section 7.3.2.2 has to be used. This combination of sampling and a small solution space can make it hard to even find a feasible (desirable) sequence. Especially, if many constraints of the very restrictive type 7.4 are used (as no conflicting trains are allowed between s_a and s_b, which can be a large proportion of the trains considered).

By searching instead of sampling, the probability of finding a feasible solution in a very small solution space increases. However, by searching, sequences are not chosen randomly anymore and the found sequences might be biased as some sequences are found more easily. This conflicts with the requirement that the produced sequences must be random to produce a result that is as close as possible to the true one (where all permutations are considered). However, as this thesis is focused on strategic planning, output including a marginal amount of undesirable sequences is preferred to no output at all.

7.3.3 Calculation of infrastructure occupation

To calculate the infrastructure occupation, a 2-step approach is used. First network headway times, $h_{ij} \in H$, are calculated between all l^2 pairs (2-tuples) of train types based on headway times defined on each edge used by both runs, H_e. To calculate the minimum network headway between a pair of train types the algorithm given in algorithm 5 (appendix B.2) is used. This algorithm is an asynchronous scheduler based on the first-in first-out principle (FIFO) and has been developed by this thesis. Essentially, for the second train in a pair the bounding link (bottleneck) is found along the run's edge sequence (route). The headway to the first train on each (common) edge is then adjusted to the earliest arrival time based on the bounding edge (bottleneck). Thus algorithm 5 gives

the required minimum network headway between a pair of train types as the difference between the earliest departure time of the two trains. Note that the minimum network headway may be negative as train types may not necessarily enter the network in the same place. Algorithm 5 has a computational complexity of $O(|E| \cdot |S|)$.

7.3.3.1 Acyclic operation

Given the network headways ($h_{ij} \in H$) and the current sequence considered, Z, the problem of finding the infrastructure occupation of the sequence is formulated as a longest path problem in a graph representing the sequence and minimum headway times.

In this minimum headway time graph (from now on shortened headway graph), $G_h = (V_h, E_h)$, each train, z, in the sequence Z is a vertex. For each vertex (train) in the headway graph, z_i, there are directed edges, $e_{z_i z_j}$, to all subsequent vertices (trains) in the sequence, $(z_j : j > i)$, whenever there is a relation (train i and j share part of the network). Each edge is weighted by the minimum network headway time, h_{ij}, between train i and j, calculated in the first step.

Finding the longest path in the graph ensures that all minimum headway times are met. Thus a timetable derived from the longest path will be feasible as edges in the longest path are the most restrictive (longest) minimum headway times between trains in the sequence. Furthermore, as the edges in the graph considered are weighted by the minimum network headway times only, the longest path is not longer than absolutely necessary to meet feasibility (minimum headway times) and it therefore corresponds to a compressed timetable with no critical buffer times (see section 6.4.2). If critical buffer times would be present in a compressed timetable based on the longest path found, it would imply that edges are weighted more than the minimum headway, which is not the case.

The infrastructure occupation rate can subsequently be derived by the ratio between the length of the longest path (infrastructure occupation) and the time period considered:

$$\text{infrastructure occupation rate} = \frac{\mathit{length}(\text{longest path in headway graph})}{\text{Time period}} \tag{7.7}$$

Furthermore, a compressed timetable can be derived from the longest path to all vertices (trains). In figure 7.3 an example of a headway graph, G_h, with four trains is shown. In the figure trains 1, 3, and 4 compose the longest path and thus the length of this path is the infrastructure occupation. Furthermore, it can be observed in the figure

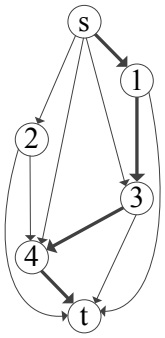


Figure 7.3: Headway graph for a sequence of four trains. Longest connected path Z'_c marked in bold given equal weights on edges.

that there are no edges between train 2 and trains 1 and 3, which represents the lack of a conflict between these trains (they do not use the same part of the network).

The longest path problem is NP-hard in general graphs. However, as the graph constructed is a directed acyclic graph the problem can be converted to a shortest path problem with negative weights for which there exist linear time algorithms (Sedgewick, Robert, 2011). The longest path is found in $O(|V_h| + |E_h|)$ time by visiting the vertices of the headway graph in topological order (already given by the sequence). For each vertex, the longest path ending at the vertex is found using dynamic programming. See appendix B.3 for the longest path search algorithm.

The problem of finding the longest path in a directed acyclic graph is the same as finding the critical path in project planning, thus the longest path may also be denoted as the critical path in the headway graph of the given sequence. Finding the infrastructure occupation is thus analogous to finding the project duration (length of the critical path) where each activity duration corresponds to the minimum network headway time between two trains in the sequence.

7.3.3.2 Cyclic operation

Minimum headway time graph For a cyclic timetable, all train runs repeat in a cyclic pattern. To compute the infrastructure occupation, it is not sufficient to look only at one repetition. In the cyclic case, the interaction between trains from one sequence repetition (cycle) to the next has to be accounted for.

In figure 7.4, four repetitions of the same train sequence are shown, each repetition is shown in a different color. The edges represent minimum headway times. In principle there are edges (headways) between each pair of vertices. However, in figure 7.4, only edges that lie on a longest path from the source (s) to a certain vertex are shown.

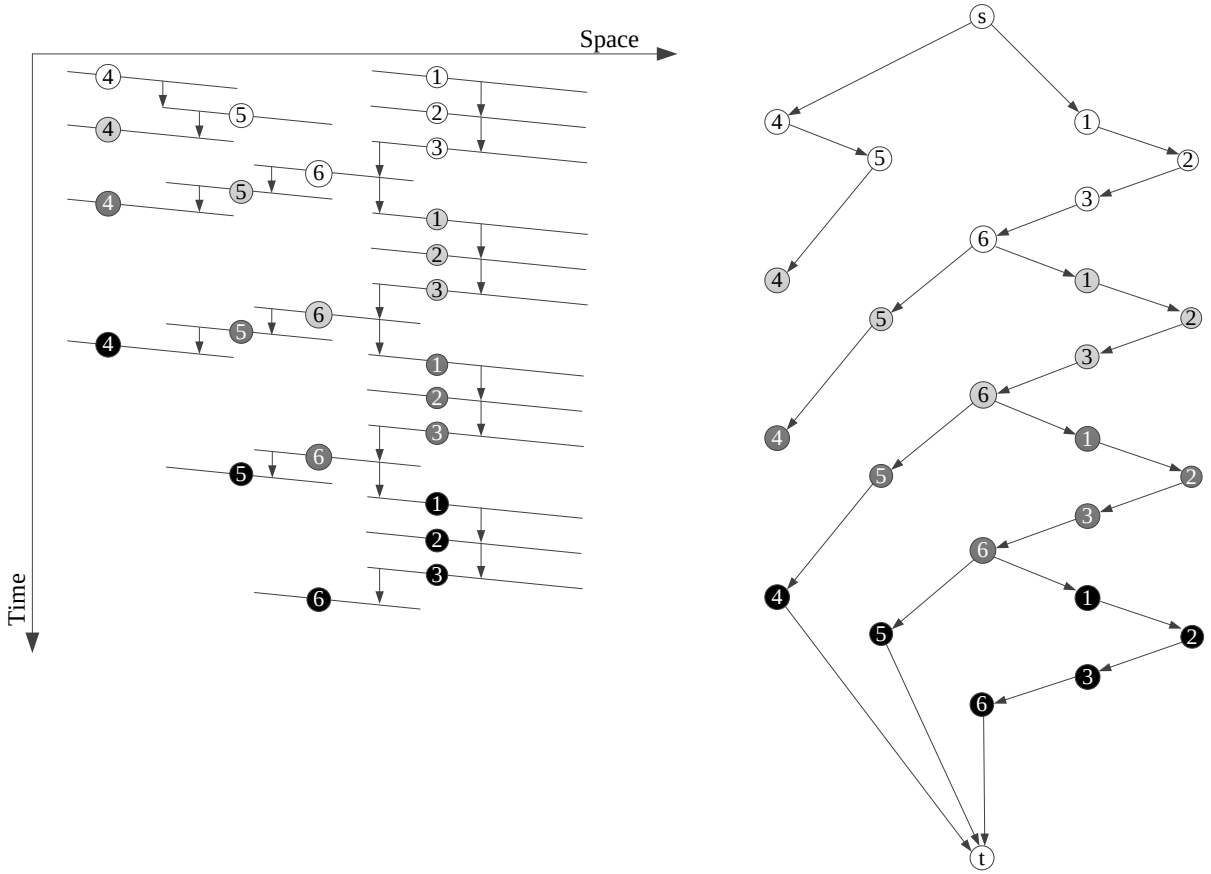


Figure 7.4: Graphical timetable (left) and headway graph (right) for a sequence of seven trains. Only edges that are part of the longest path from s to a vertex (train) are shown. Colours denote different repetitions of the sequence.

For any given vertex (train), z, in any expanded headway graph (with repetitions of the train sequence added), only edges (headways) originating from the previous $|Z|$ vertices (trains) in the topological order when searching for the longest path needs to be considered. That is, only minimum headway time requirements going one train sequence repetition (cycle) back has to be considered. Edges going more than $|Z|$ vertices back will not be part of the longest path and are thus not necessary. This can be observed by contradiction:

Consider an arbitrary vertex (train), z_x^a, in repetition x of train type a with minimum headway time h_{aa} between trains of type a. If z_x^a is part of the longest path, then it must hold that the path $z_{x-2}^a \rightarrow z_{x-1}^a \rightarrow z_x^a$ is shorter than the path $z_{x-2}^a \rightarrow z_x^a$. That is the edge from vertex z in repetition $x-2$ to z in repetition x is part of the only longest path. This cannot be true as seen by the false inequality $h_{aa} + h_{aa} < h_{aa}$ as $h_{aa} > 0$.

Similarly, with z_x^a and a preceding train of another type b, z_x^b, in the longest path then again it must hold that the path $z_{x-2}^b \rightarrow z_{x-1}^a \rightarrow z_{x-1}^b \rightarrow z_x^a$ is shorter than $z_{x-2}^b \rightarrow z_x^a$, which leads to the false inequality $h_{ba} + h_{ab} + h_{ba} < h_{ba}$ as $h_{ba} + h_{ab} > 0$.

This property significantly reduces the number of edges in the graph and thereby the running time of the algorithm as fewer edges have to be considered in the longest path search for a cyclic operation pattern.

The necessary number of repetitions to find the longest path Although, only dependencies between two sequence repetitions have to be considered, it does not guarantee that two repetitions is enough. This is caused by the fact that this model is considering networks and thus not all trains have direct interaction with each other. Two repetitions is only sufficient in the case where a part of the network is considered where all trains have direct interaction.

The longest path algorithm traverses the headway graph in topological order and records the longest path to each vertex (train) in the order. A longest path in one repetition of the train sequence (cycle) may change in the next repetition (cycle) as interdependencies between trains from early repetitions affect later repetitions. Two special cases have been observed where more than two repetitions are necessary:

Special case 1: In the example in figure 7.4, the longest path $4(white) \rightarrow 5(white) \rightarrow 4(lightgrey)$ initially constitutes the longest path for train 4 (light grey). However, in subsequent repetitions of the sequence the longest path changes as dependencies from train 1 to 3 affect train 4–6. Consequently, the longest path to all vertices does not settle until the third repetition (dark grey). If even more train types were added on the left side of the graph, in the same manner as train 4–6 (figure 7.4), even more repetitions would be needed to obtain the true longest path. However, from figure 7.4 it can be seen that if the sequence was $(1, 2, 3, 6, 5, 4)$ instead of $(1, 2, 3, 4, 5, 6)$ the longest path to all vertices would be found in the first repetition.

Special case 2: In addition to the case described above, a longest path may form across two or more sequence repetitions. E.g. the longest path may start with train (vertex) z in repetition x, in repetition $x + 1$, z is not part of the longest part, however it is again part of the longest path in repetition $x + 2$. Thus the longest path forms a "loop" (down the graph) from repetition x to $x + 2$. This is not desirable as the goal is to find a longest path that repeats in each repetition. That is train z should be part of the longest path in both repetition x and $x + 1$.

Based on the two special cases for cyclic timetables described above, the longest path search needs to be extended with the following to obtain the infrastructure occupation (minimum cycle time) for a cyclic train sequence:

1. The sequence repetitions (cycles) needed to obtain the longest path depend on train and network interdependencies.

2. The longest path found may be constituted of trains (vertices) across more than one sequence repetition (cycle)

Adjusted longest path search The first special case is addressed by testing that the longest path derived by the longest path algorithm (see appendix B.3) is the same from repetition to repetition. As the risk of failing this test is highest for the first couple of repetitions (as seen in the example in figure 7.4), the check is done after a start threshold of u repetitions. If the check fails, the longest path has not been found or the longest path has formed across several repetitions according to special case 2.

The search is continued, and if the longest path to all vertices does not change, it is assumed that special case 2 holds. The minimum cycle time (infrastructure occupation) is then derived as the maximum path length between any vertex (train) z_{x-1} (repetition $x-1$) and z_x (repetition x) where either z_{x-1} or z_x is part of the longest path. To obtain the compressed (cyclic) timetable the weight (minimum headway time), h_{ii}, on all edges e_{ii} are changed to the minimum cycle time obtained and the search is continued until a longest path to all edges is found that repeats in consecutive repetitions.

The adjusted algorithm is given in algorithm 1 and has linear computation time in the order of the number of trains. It should be noted that it is only necessary to keep two repetitions at all times as edges does not span across more than one repetition as described earlier.

Algorithm 1 does not guarantee that a correct longest path is found. However, experimental tests on the network given in section 7.4 show that only 0.12% of sequences is of special case 1 after 3 repetitions and 0.2 % of sequences of special case 2 for which all are captured by the algorithm with $u = 5$.

In practice, as there are so few of these sequences, the infrastructure occupation values estimated, based on these sequences, would only constitute a very small amount of the total set of values calculated. However, if one has to be certain that the correct longest path is found, u can be set to the number of cycles (sequence repetitions) per day. This will ensure that the last cycle, u, of the day is correct and it can therefore be copied to the previous $u-1$ cycles.

Algorithm 1: ALPS: Adjusted longest path search for cyclic train patterns

Data: Minimum headway graph $G_h = (V_h, E_h)$, check start u, and train sequence Z (topological sort of G_h)
Result: Longest path (LP) to all vertices, Z', in G_h, longest connected path, Z'_c, and start times, $t_z \in T_z$

```
Do LPS (Longest Path Search) on G_h
loop ← 1
while checks not passed do
    Add sequence repetition to G_h
    Do LPS on added repetition
    loop ← loop + 1
    if loop > u then
        if longest path match between repetition u and u − 1 then
            return Z', Z'_c and T_z from last LPS execution (u)
        end
    end
    if loop > u + 2 then
        if no change of LP to all vertices from u to u − 2 & u − 1 to u − 3 then
            /* Find max time between same train in repetition u and u − 1
               */
            t_max ← max_z(z_u → z_{u−1})
            h_ii ← t_max   ∀i ∈ V_h
            Add repetitions and do LPS on added repetitions until longest path
             match between repetitions
            return Z', Z'_c and T_z from last LPS execution
        end
    end
    if loop = max_loops then
        Print error and return
    end
end
```

7.3.4 Calculation of capacity consumption using stochastic simulation

To derive the capacity consumption, critical buffer times between trains have to be estimated to obtain results that are robust against delays (see section 3.1).

To do this an event-driven stochastic simulation procedure has been developed by this thesis. For each desirable sequence considered a compressed timetable is generated based on the longest path derived using the procedure described in section 7.3.3.

Based on the compressed timetable, the simulation iterates a predefined number of times where random initial delays are sampled and added to train runs at specified network locations (e.g. a station). Only smaller initial delays are considered and include extended dwell and running times due to e.g. smaller failures or slow passenger exchange. The initial delays are sampled from one or more probability distributions.

Running time supplements are added to the minimum running times (the fastest possible running time) during the planning process. These supplements are put in to absorb smaller delays of a train. Thus the running time of a train, s, consist of the minimum possible running time plus a supplement, $T_s^{'}$.

To determine how much of the supplement that can be used to absorb a delay in the simulation, $y_{s,e}^{l}$ is defined as the relative position of an initial delay l on an edge e of train run s.

It is possible to sample more delays, $l \in L$, for a run, s, on an edge, e, with different values, $d_{s,e}^{l}$ and $y_{s,e}^{l}$, which will then be aggregated according to:

$$d_{s,e}^{\sigma} = \sum_{l \in L} d_{s,e}^{l} \tag{7.8}$$

$$y_{s,e}^{\sigma} = \sum_{l \in L} \frac{y_{s,e}^{l}}{d_{s,e}^{l} \cdot (1/d_{s,e}^{\sigma})} \tag{7.9}$$

Where $d_{s,e}^{\sigma}$ respectively $y_{s,e}^{\sigma}$ is the aggregated delay and position on a edge for a specific train run. This aggregation is useful as it produces one initial delay from multiple sampled delays for an edge. For instance when initial delays have to be sampled for multiple stops (dwell time extensions) on a single edge.

Given the compressed timetable and the initial delay, the delay propagation can be derived. For instance given an arbitrary train run on an edge, the following procedure is done. Based on the aggregated delay, $d_{s,e}^{\sigma}$, and its position, $y_{s,e}^{\sigma}$, the amount of running time supplement available for the train on the edge is calculated. Given this supplement, the initial delay, and delay from earlier edges or trains, the amount of delay of the train

at the end of the edge can be calculated. If there is a delay, this causes a shift in the scheduled time that may produce a conflict with the following train. If a conflict arises, delay propagates from one train to the next.

The described procedure is done for all trains in the compressed timetable. This yields a "realised" compressed timetable, where delay propagated means that the sequence considered takes longer to complete than without delays (the infrastructure occupation).

This prolongation of time is an indication of where there should have been added buffer time to neutralize the delay propagation. Thus the capacity consumption of the sequence (based on the delay sampled) is the time it takes to complete the sequence with delays.

As the initial delays are random, the procedure has to be repeated a sufficient number of times (iterations) to yield an accurate result. In practice this lead to a series of values ranging from the best case, where sampled initial delay causes little delay propagation, to the worst case, where sampled initial delay causes a high delay propagation.

The above procedure estimates the capacity consumption for one compressed timetable based on a generated sequence. This procedure thus has to be repeated for all the generated sequences as described in section 7.2. The simulation procedure has a worst computational time complexity of $O(|S| \cdot |E|)$ per iteration and is thus fast and efficient.

As a final remark, it should however be noted that the approach used is not as precise as a microscopic simulation. Yet, it is considerably faster due to the event-driven approach and the mesoscopic infrastructure model used.

7.4 Case study

In this section, the developed framework and model is applied to a network in Southwestern Denmark consisting of 161 kilometres of double track, at-grade and out-of-grade junctions and four terminal stations. Over a period of one hour, a total of 18 trains should enter the network distributed over 2 express trains, 6 intercity trains, 6 regional trains and 4 freight trains according to the plan of operation. The network is depicted schematically in figure 7.5 with tracks used in normal operation, while table 7.2 shows the routes used.

7.4.1 Probability distributions for initial delays

Similar to Corman et al. (2011), the perturbed scenarios are constructed for the stochastic simulation by sampling initial delays from a given probability distribution when the trains enter the network at entrance stations OD, FA, VM and ES and at each stop in the

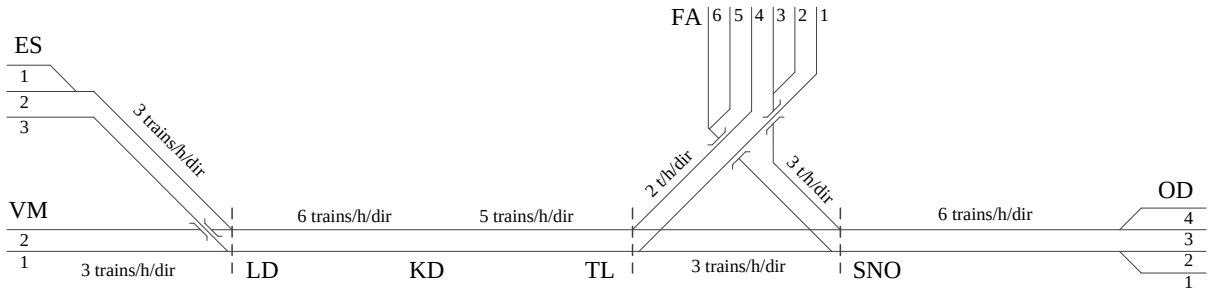


Figure 7.5: Case network in Southwestern Denmark.

network. Yuan (2006) finds that the Weibull distribution is the best distribution model in general for non-negative arrival, departure and dwell delays. This probability distribution is therefore used to sample from. The distribution parameters used for this case are based on estimated parameters from the work of Corman et al. (2011); Jensen and Landex (2013a); Yuan (2006) for a similar service pattern of regional and intercity trains at The Hague CS, Utrecht CS and Skanderborg (in Denmark) stations respectively.

Only a maximum delay of 10 minutes is considered corresponding to the smaller delays occurring in normal operation. Sampled delays are therefore truncated at 10 minutes, which is also done by Kroon et al. (2008b) in their stochastic optimisation. Dwell time extensions are modelled by two shifted Weibull distributions. One with shape 2.0, scale 75 and shift parameter -30 for stops with a planned dwell time of 60 seconds. And the other with shape 2.0, scale 60 and shift parameter -30 for stops with a planned dwell time of 30 seconds. Dwell time extensions are added for each stop of a train. For some trains in the case, this means that delays are added multiple times on an edge corresponding to the number of stops.

The running time supplement is set to 9% for all trains corresponding to the recommendation given by Rail Net Denmark in the interval 141-160 kph (Schittenhelm, 2011).

7.4.2 Plan of operation

Table 7.2 shows an overview of the trains that should traverse the network together with the route and Weibull distribution parameters for the input delays when a train enters the network. Train IC-B1 and IC-B2 is coupled between OD and KD. At KD, IC-B1 and IC-B2 depart to/arrive from ES and VM respectively. This is modelled using an inclusion constraint as described in section 7.3.2.3.

The most occupied section is the line between OD and SNO with 6 trains/hour per direction. This is also the part of the network with the highest heterogeneity in running times. Between OD and SNO the slowest train (RE-A) is 18 minutes slower than the

Table 7.2: Train types in network with route used, number of trains per hour and Weibull distribution parameters for input delay at the entrance to the network.

Train	Route	#/h	Shape	Scale	Running time
Express train	1: OD(4) → FA(2)	1	0.9	70	31 minutes
↪	2: FA(6) → OD(2)	1	0.9	80	31 minutes
IC-A	1: OD(4) → FA(2)	1	0.9	60	33 minutes
↪	2: FA(6) → OD(2)	1	1	65	33 minutes
IC-B1	3: OD(3) → ES(1)	1	0.9	60	67 minutes
↪	4: ES(3) → OD(1)	1	1.1	55	67 minutes
IC-B2	31: KD(2) → VM(2)	1	–	–	50 minutes
↪	42: VM(1) → KD(1)	1	0.9	60	50 minutes
RE-A	5: OD(4) → FA(3)	1	1.1	45	48 minutes
↪	6: FA(5) → OD(1)	1	1.1	45	48 minutes
RE-B	7: FA(4) → ES(2)	2	1	70	64 minutes
↪	8: ES(3) → FA(1)	2	1.1	45	64 minutes
Freight	9: OD(3) → VM(2)	2	1.4	200	72 minutes
↪	10: VM(1) → OD(1)	2	1.4	200	72 minutes

express train that uses 26 minutes to travel this distance and 4 minutes slower than the freight train, which uses 40 minutes. The second most occupied section is between TL and LD with 5 trains/hour per direction TL-KD and 6 trains/hour per direction KD-TL. However, the heterogeneity in this section is lower. Therefore, the bottlenecks in the network are expected to be the section between OD and SNO, and in the junctions SNO and TL, as the minimum (block) headway time is the same in the whole network (estimated 150s). Exclusion constraints are added for RE-B and IC-A/IC-B train types. Thus sequences where two trains of type RE-B follow each other or sequences where trains of type IC-A and IC-B (or vice versa) follow each other are excluded.

7.4.3 Computational parameters

The model described in section 7.3 has been implemented in C++. Furthermore, multi-threading is used in the stochastic simulation to enable the use of all CPU threads for faster computation by assigning different iterations to each of the threads. The model is executed on a Windows 7 laptop with an Intel Core i7-2640M processor (dual-core/four threads) and 4 GB of RAM. The stochastic simulation is executed 60 times with different samples of input delay. 3 warm-up cycles (with different initial delay samples) are completed to ensure that the simulation is in a realistic state before measuring delay propagation. As 18 trains (14 train types) are scheduled to enter the network, there are $(18-1)!/(2!^4 \cdot 1!^{10}) = 22.2 \cdot 10^{12}$ possible permutations of train sequences, and sampling is thus needed to obtain

the output distributions of the infrastructure occupation and capacity consumption. For convergence in the sampling, the sampling is stopped when the mean respectively standard deviation of the output distribution is within a tolerance, ϵ, of 0.1% for five consecutive measurements done at an interval of 100 samples.

7.4.4 Tests and scenarios

In the following subsections, the infrastructure is assessed with the plan of operation just described using the developed model. First some tests are done to show the impact of using exclusion constraints and how divisions of the network into routes and line sections affect the output distribution. Secondly four infrastructure scenarios are considered: a base scenario as depicted in figure 7.5 and three upgrade scenarios where TL, SNO, and both TL and SNO, respectively, are upgraded from at-grade junctions to out-of-grade junctions by adding a fly-over. Thus crossing conflicts between opposing directions are eliminated as trains passing the opposing direction can use the fly-over.

7.4.5 Experimental results

The cumulative distribution of the infrastructure occupation for the network and route partitions has been calculated using the model without the stochastic simulation. The computation time for the network is approx. 1-2 seconds using one thread of the CPU when exclusions are not considered and approx. 10 seconds when they are. The convergence criteria is met after approximately 40,000 samples. For the route and line partitions, the computation time is less than a second. Sampling is used for route 3 and 9, while all unique permutations are assessed for the remaining routes as the total number of permutations is low. For route 3 and 9 convergence is achieved after approximately 31,500 samples.

Figure 7.6 shows the cumulative distribution of the infrastructure occupation in the base network with partitions containing the whole network, each route, and line sections between OD and ES, respectively. For the network results both results with and without the sequence exclusions are shown. With the exclusions the network infrastructure occupation ranges from 60.8% to 203.3 % with a median of 105.4 % as seen from figure 7.6 (maximum not shown). Without exclusions the occupation rate naturally decreases as constraints are removed. This results in a reduction of the median to 99.9 %.

From figure 7.6 it can be seen that the distribution of the infrastructure occupation is different when only the edges used by the different routes are considered. This is expected as the constraints are changed. If they were to be the same for a route and the whole

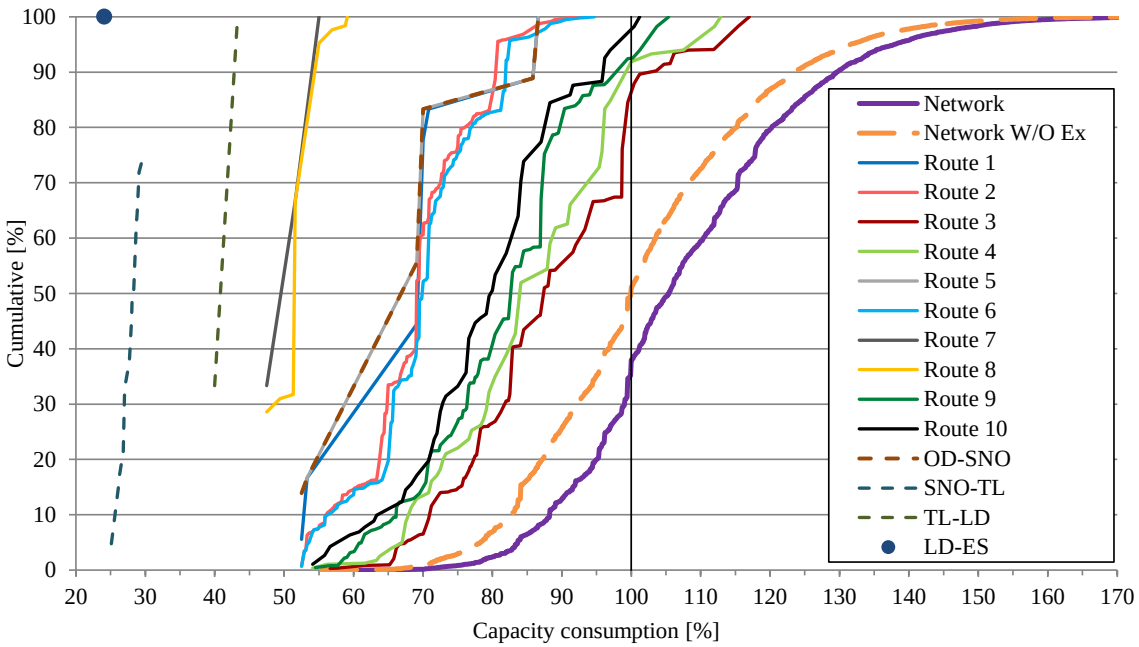


Figure 7.6: Base scenario results for the network with and without exclusions, for routes, and for four line sections between OD and ES (route 3).

network, the bottleneck(s) of the network would always be one or more of the edges in the route considered. Furthermore, it can be seen from the figure that when dividing route 3 (the most trafficked route) into line sections, according to UIC 406 recommendations, the infrastructure occupation decrease even more. Using the line sections to conduct an analysis on the operation between SNO and ES would thus lead to the false conclusion that all trains sequences have an infrastructure occupation below 100%. The results for route 3 show that this is incorrect.

This partitioning of the network illustrates that the result depends on the division of the network. As mentioned in the chapter introduction, this is one of the problems with decomposition of the network for capacity analysis. Larger analysis areas will lead to higher figures of infrastructure occupation.

Besides illustrating the point of network effects in capacity analyses, the results depicted in figure 7.6 also show the unused capacity on the routes. For example routes 7 and 8 used by RE-B trains have a lot of unused capacity. This is expected as they operate on a part of the network where the traffic intensity is lower than in other parts of the network and the traffic is more homogeneous.

Likewise, the routes that include junction TL (that has a high amount of traffic) have less unused capacity. Routes 3, 4, 9 and 10 that span the largest part of the network have the least amount of unused capacity. Of these routes, route 3 and 9 have the least unused

capacity compared to the opposite direction routes, 4 and 10 respectively. Route 3 and 9 pass both at-grade junctions (TL and SNO), while route 4 and 10 do not. Thus, as expected it appears that the at-grade junctions posses a capacity constraint. Regarding the unused capacity, it is important to note that in most cases not all the unused capacity can be used. This capacity can only be utilised, if the buffer times are large enough to hold a relevant train path on the route. Remaining capacity will be lost, for more information, see UIC (2013).

7.4.6 Scenario results

In the previous section it was concluded that the at-grade junctions pose a capacity constraint. In this section, an assessment of the capacity consumption is conducted where the junctions are upgraded to out-of-grade junctions based on the input delays defined in table 7.2. Running times for the affected trains are kept the same.

Executing the stochastic simulation over 60 iterations yields the results depicted in figure 7.7. Using all four threads/two cores of the processor, the computation time per iteration is approximately 13 seconds. In total the computation time for the whole network is approximately 900 seconds (15 minutes) for all 60 iterations. In this timespan approximately 2.4 million desirable compressed timetables are generated and simulated.

Figure 7.7 shows minimum, 25th percentile, median, 75th percentile and the maximum capacity consumption distributions over all iterations for the base case and the upgrade scenarios. From the results of the base case, it can be observed that a significant amount of the distribution is above 100 %. In fact only 7% (median) of the train sequences have a capacity consumption of maximum 100 %. Thus, if the system were to be run at random according to the plan of operation, only 7% of the possible train sequences will be robust enough to cope with delays occurring every second day.

Comparing infrastructure occupation and capacity consumption, the infrastructure occupation is 83.7 % at cumulative 5.2% (75th percentile) and 85.8% at 7% (median) (see figure 7.7 (top)). This is higher than the recommendation from the UIC (2013) of 75% for mixed traffic lines in the rush hour. However, this threshold is rather coarse as also described in section 7.1. Furthermore, the threshold is based on smaller line divisions (which cause lower values).

For the upgrade of TL, figure 7.7 shows that the maximum distribution is more extreme than for the other scenarios. This is a consequence of the stochastic simulation, as the results contain iterations where the amount of input delay is sampled at critical places yielding a high delay propagation. thus resulting in the maximum distribution seen.

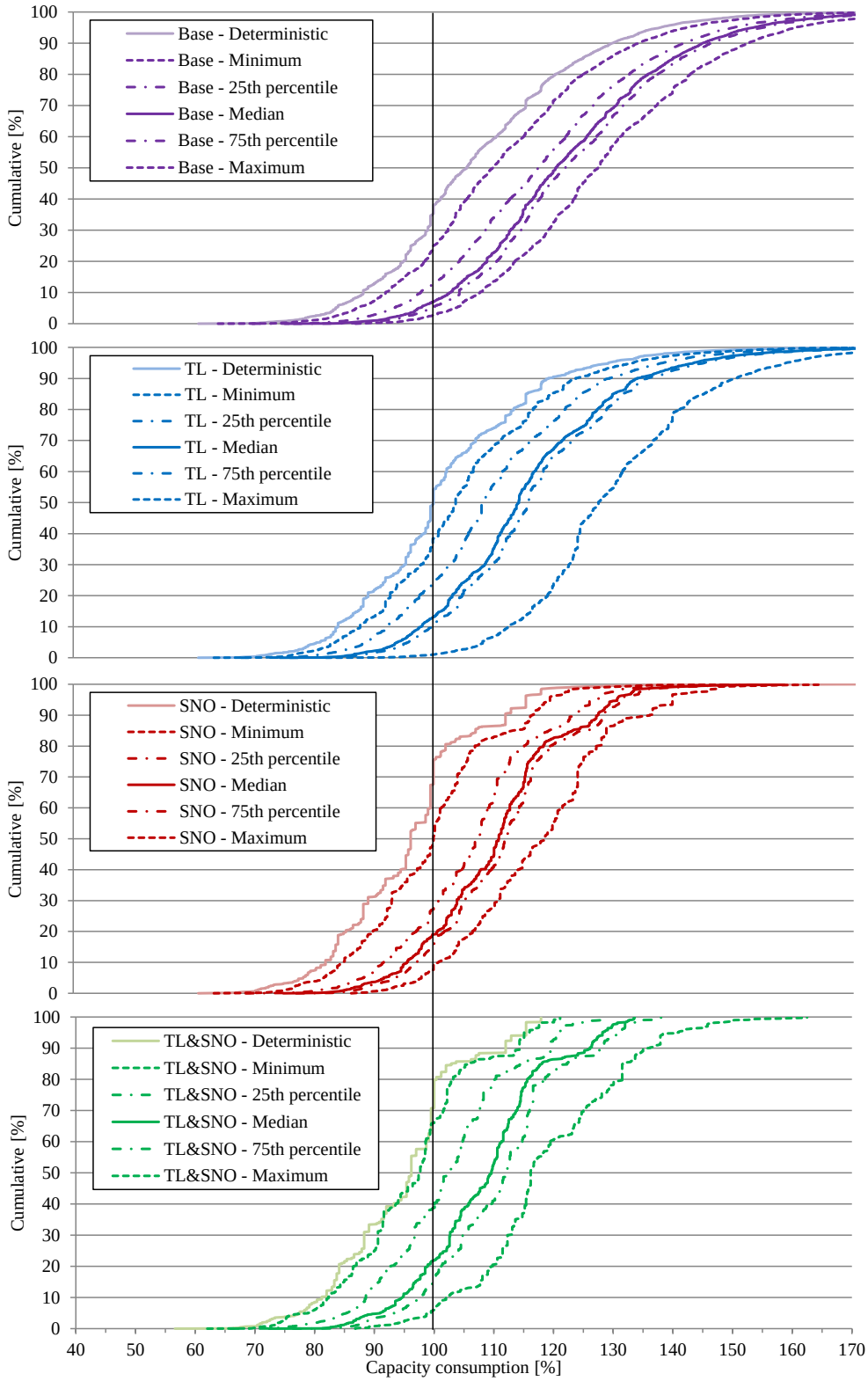


Figure 7.7: Minimum, 25th percentile, median, 75th percentile and maximum cumulative capacity distribution as well as infrastructure occupation for all scenarios (base: no upgrade, TL: upgrade of junction TL, SNO: upgrade of junction SNO and TL&SNO: upgrade of both junction TL and SNO). See figure 7.5 for location of junctions.

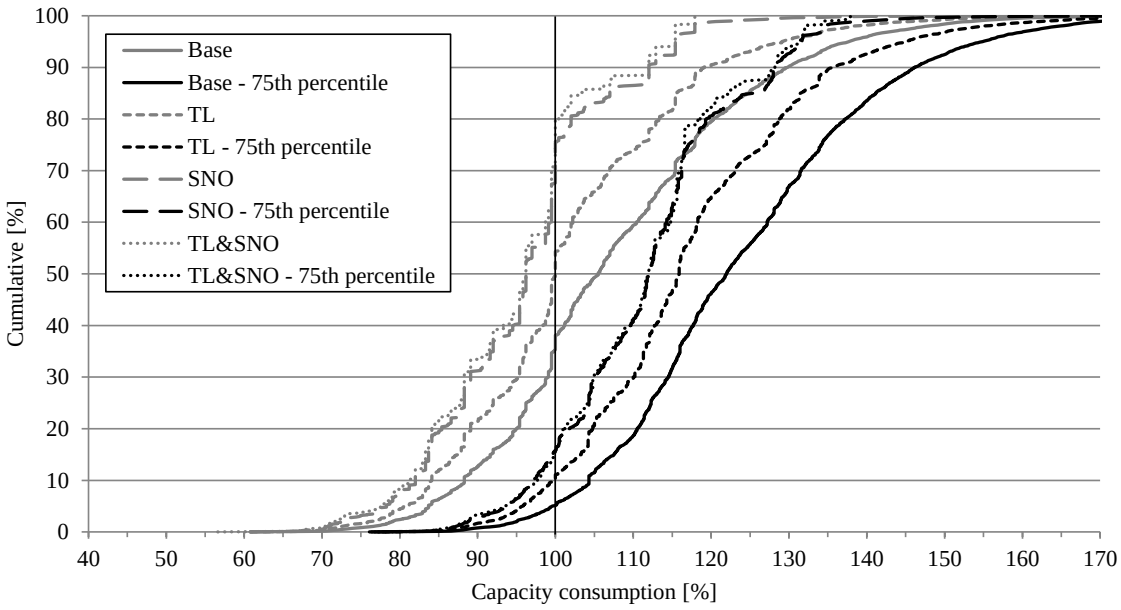


Figure 7.8: Scenario results for upgrading junctions.

By upgrading the junction(s) the infrastructure occupation and the capacity consumption can be lowered significantly as seen in figures 7.7 and 7.8. This decrease can be used to run more trains or maintain the same amount of trains with an additional robustness. The additional robustness gained is also expressed by the fact that the amount of consecutive delays (risk of delay propagation) is reduced. These consecutive delays have been derived from the stochastic simulation model developed.

From the results in figures 7.7 and 7.8, it can be observed that the upgrade of SNO produces a higher capacity gain than the upgrade of TL. It can furthermore be observed that the amount of feasible train sequences is increased with the upgrades. With the TL upgrade scenario yielding a two-fold increase, while the feasible amount of sequences in the SNO upgrade is almost tripled. An upgrade of both junctions will triple the amount of feasible train sequences in the system. However, it can also be observed that upgrading both junctions provides only a little benefit compared to only upgrading SNO. Thus the upgrade of SNO should be prioritised over the upgrade of TL.

7.5 Summary and conclusions

This chapter has presented a framework for the assessment of infrastructure occupation and capacity consumption in railway networks in the strategic planning process. The framework outlines the overall procedure which may be implemented in any way found suitable. In the framework, a predefined timetable is not needed as input, only the plan

of operation (services intentions) in the form of the number and type of trains including running and headway times as well as the infrastructure layout. Based on this different train sequences (train orderings) are generated and evaluated. The output is thus a series of capacity consumption values (a distribution) rather than a single value.

Furthermore, in this chapter, an adjusted longest path search based on the minimum headway times required between trains was developed to derive infrastructure occupation values. To obtain the capacity consumption, this thesis developed and used a stochastic simulation to estimate buffer times (to account for robustness against delays) based on the delay propagation caused by a set of input delays.

The developed framework, and the model implementation of it, was used on a case network in Southwestern Denmark with two at-grade junctions where four scenarios were considered: a base scenario and three scenarios where the at-grade junctions were upgraded to out-of-grade junctions. In all scenarios, the model can quantity the improvements in network infrastructure occupation and capacity consumption.

Furthermore, results from the case show that the results of capacity assessments are very much dependent on the size of the network considered. This is a consequence of adding more constraints to the calculation as the network grows. Therefore, this thesis suggests considering the whole network and in this context to evaluate capacity consumption rather than infrastructure occupation to also include the aspect of robustness.

8 Capacity determination framework

This chapter is an edited version of the following article:

L. W. Jensen. An optimisation framework for determination of capacity in railway networks. In *Proceedings of Conference on Advanced Systems in Public Transport (CASPT)*, Rotterdam, 2015

In this chapter, an optimisation framework to estimate the capacity in a railway network is developed. In this context, the capacity is defined as the number of trains of each train type that is able to traverse the network under a predefined mix of train types plus trains that additionally can be added. The solution is subject to the given infrastructure, the train types, characteristics of the train types, a desired mix of train types, and the routes used within a defined infrastructure occupation threshold, C_{max}, (to account for robustness).

The developed optimisation framework makes it possible to estimate both the maximum and practical capacity (as defined in section 6.1) by adjusting the threshold (C_{max}). In this chapter, the focus will mainly be on practical capacity. The developed optimisation framework does not require a timetable and is thus suitable for the strategic level to evaluate infrastructure alternatives and provide capacity estimates for e.g. line planning.

The chapter outline of the chapter is as follows. Section 8.1 describes directly related approaches in detail before the methodology is presented in section 8.2. In section 8.3, the network of section 7.4, is used to carry out computational experiments as well as present capacity estimates. Finally, section 8.4 summarises and draws conclusions of the chapter.

8.1 Previous approaches

As described in chapter 6, there already exist several methods to analyse railway capacity. In the following, only approaches directly related to absolute capacity determination are described. See chapter 6 for more methods related to capacity assessment.

As briefly described in section 6.5.3.1, the work of Burdett and Kozan (2006); de Kort et al. (2003); Mussone and Wolfler Calvo (2013), deal explicitly with the determination of capacity in railways. These approaches are suitable for strategic planning as they require no timetable.

de Kort et al. (2003) use a probabilistic approach using max-plus algebra to determine the capacity of railway infrastructure. The approach is based on the bottleneck approach which determines the critical section (bottleneck) that limits the capacity of the system considered. The approach does not explicitly account for different train types. However, it is possible to implement this by using the probability that a train will be of a given type.

Inspired by the work of de Kort et al. (2003), Mussone and Wolfler Calvo (2013) present an optimisation framework to maximize the total number of trains in the railway system based on train conflict probabilities, where different train types can be considered. de Kort et al. (2003); Mussone and Wolfler Calvo (2013) are able to include knock-on (consecutive) delays in the railway system in a simple way.

Lastly, Burdett and Kozan (2006) describe and discuss terms for absolute capacity and utilisation levels based on sectional running times of train types, dwell times, and proportional mix of train types. Based on these, they propose an optimisation model to maximize the number of trains in a railway network. The model proposed is non-linear and is therefore only solved to a local maximum (Burdett and Kozan, 2006). The approach by Burdett and Kozan (2006) is quite flexible and is able to capture a large amount of operational characteristics. Furthermore, the approach is able to give a lower and an upper bound for capacity.

Common for the approaches by de Kort et al. (2003); Mussone and Wolfler Calvo (2013) is that they do not depend on the exact train sequence, but rather consider a weighted average. Thus, in a network with heterogeneous operation the average absolute capacity is obtained, and this capacity might therefore be increased by bundling trains.

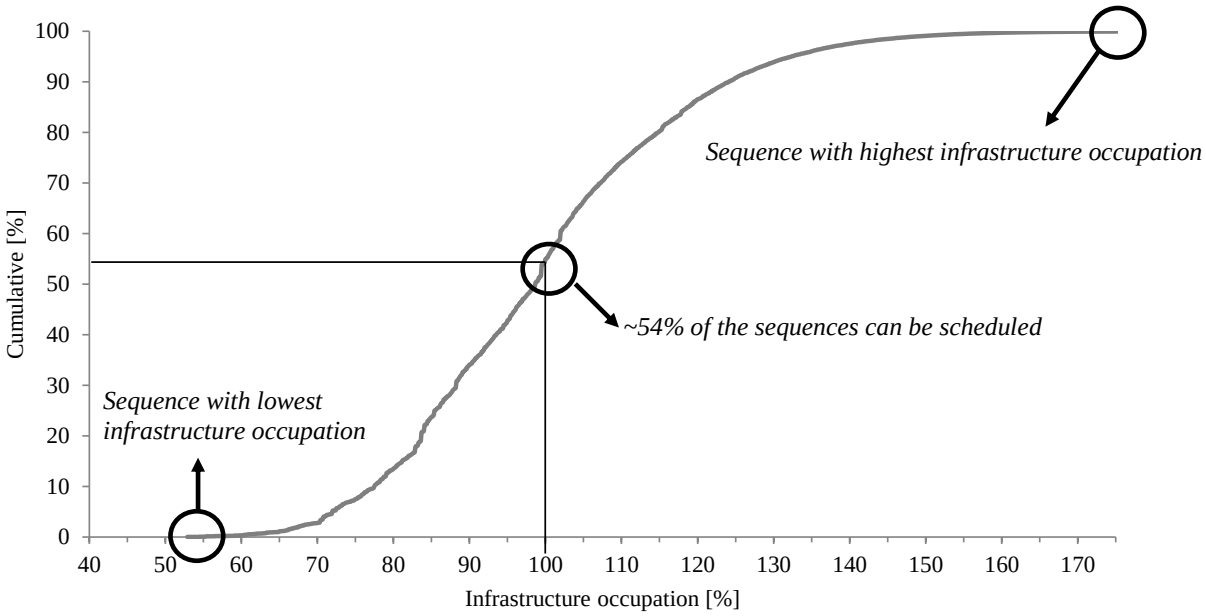


Figure 8.1: Cumulative capacity consumption distribution.

8.2 Methodology

Unlike existing methods, described in the previous section, the optimisation framework developed by this thesis is based on the model described in chapter 7. This model estimates the infrastructure occupation and capacity consumption based on a given set of trains. Using this model in the developed optimisation framework makes it possible to capture the distribution of capacity in railway networks with heterogeneous operation. I.e. the capacity can be measured relative to how many trains sequences should be feasible. The use of the model of chapter 7 is briefly described in section 8.2.1, while the optimisation framework is described in section 8.2.2.

8.2.1 Model for calculation of capacity consumption

This section briefly recaps the framework and model of chapter 7 as used in the optimisation framework of this chapter. For a more in-depth description, see chapter 7.

The optimisation framework is based on the output infrastructure occupation distribution of the capacity model 7 as shown in figure 8.1. The y-axis shows the cumulative percentage. This percentage shows how many of the possible permutations of train sequences that fit within the infrastructure occupation depicted on the x-axis. In the remainder of this chapter, the cumulative percentage is denoted as the percentile, p, while the capacity consumption is denoted by C. For a specific set a trains X the maximum

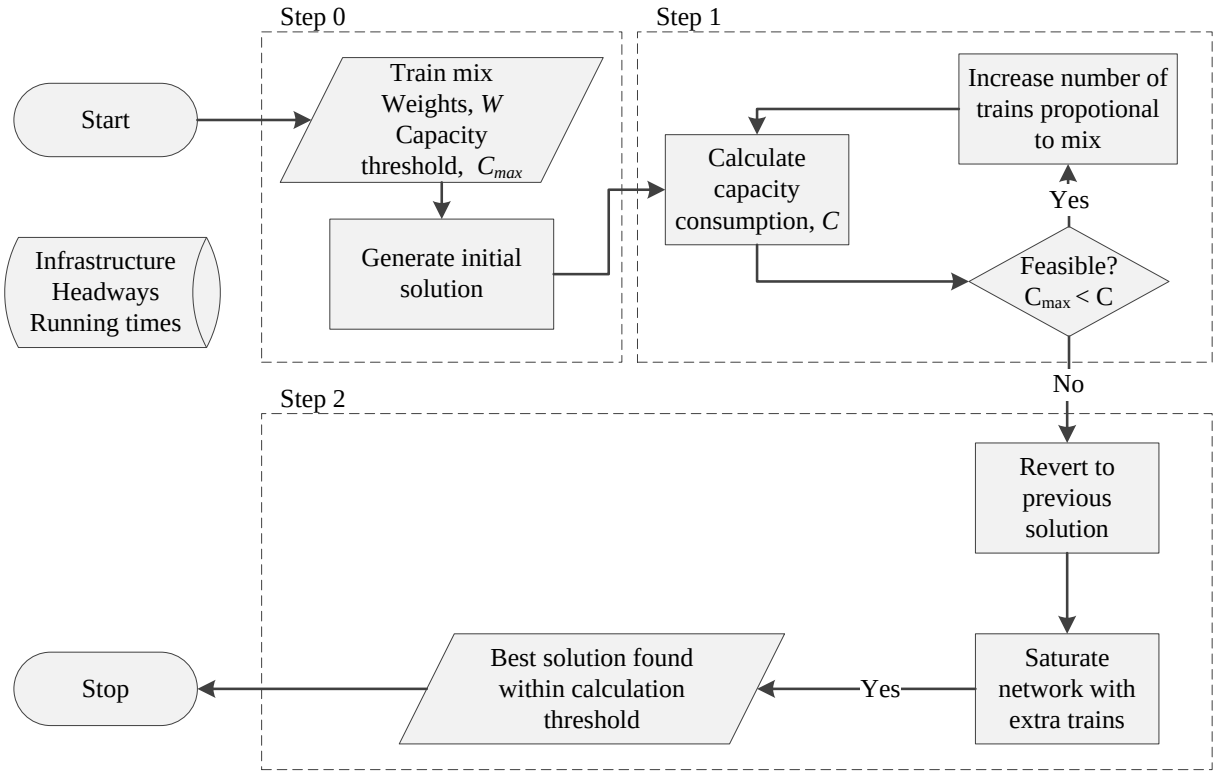


Figure 8.2: Proposed framework.

infrastructure occupation at a percentile, p, is denoted as $C(X, p)$. For instance in figure 8.1, approximately half of the analysed train sequences are feasible ($C(X, 54) = 100\%$), when neglected the need for critical buffer times. That is, the 54th percentile (close to the median) is below 100% infrastructure occupation, where sequences yielding infrastructure occupation values larger than 100% are infeasible.

For the model, the input is the number and type of trains, the network, routes as well as minimum headway times and running times. The infrastructure is represented by a mesoscopic infrastructure model that makes it possible to model double and multiple track as well as junctions. As described in section 7.2, the model is not dependent on the exact timetable as an input.

8.2.2 Optimisation framework

The optimisation framework for the determination of capacity, proposed in this chapter, is shown as a flowchart in figure 8.2. As stated in the introduction to this chapter, the capacity is defined as the number of trains respecting a given train type mix plus additional trains that the network can handle under a predefined threshold, C_{max}. A high

threshold will result in a less robust system than for lower thresholds as there will be less buffer times.

As mentioned earlier, the optimisation framework relies on the model described in chapter 7 (and section 8.2.1). However, the framework can be used with any method or model that evaluates the infrastructure occupation of a given set of trains in a given network. Although, it will not be possible to derive the capacity span, if the alternative model is not able to calculate the distribution of infrastructure occupation. If this is indeed the case, existing approaches may be more suitable (see section 8.1). If an alternative method is used for infrastructure occupation determination in the optimisation framework it must fulfill the following requirements:

- If an extra train is added to a network the infrastructure occupation will never decrease compared to the previous solution (without the extra train)
- Likewise, if a train is removed the infrastructure occupation will never increase

Not all models fulfill these requirements. For instance, models where scheduling decisions is made by slowing down fast trains and/or re-routing trains, the addition of a train might lead to a lower infrastructure occupation.

The proposed optimisation framework consists of three steps which will be described in the following subsections.

8.2.3 Step 0 and 1

Step 0: In this step, the data is loaded and an initial solution is generated. This initial solution generated is the minimum solution respecting the given mix. For instance with four train types constrained to a mix of 20-30-10-40%, the initial solution will be 2, 3, 1 and 4 trains of each type, respectively.

The infrastructure occupation threshold, C_{max}, is an important input to the framework. This will be 100% if the maximum capacity has to be obtained or less if the practical capacity has to be obtained. As defined in section 6.1, the maximum practical capacity is the maximum capacity that can be utilised if the operation has to remain stable when smaller delays occur. As also described earlier, UIC (2004b, 2013) has made recommendation for maximum infrastructure occupation for stable operation. For instance, 75% in the peak hour for heterogeneous (mixed) operation.

As described in section 8.2.1, the output from the infrastructure occupation calculation is a distribution of infrastructure occupation when train operations are heterogeneous. Thus in the optimisation a percentile, p, of this cumulative distribution has to be chosen

to obtain one single infrastructure occupation figure, C. The 0th percentile will yield the maximum number of trains (upper bound), while the 100th percentile will yield the minimum number of trains (lower bound). A natural choice would be the 50th percentile or the 100th percentile (lower bound). Due to variance in sampling sequences it is however not recommended to use the 0th and 100th percentiles. A span of capacity, as multiple of predefined mix plus additional trains, can be derived by executing the optimisation several times with different percentiles, for instance for the 5th, 50th and 95th percentiles.

Step 1: In this step, the capacity respecting the train mix is estimated. First the infrastructure occupation of the initial solution is calculated. If the infrastructure occupation calculated, C, is higher than the threshold C_{max}, it is not possible to schedule the given train mix with the threshold, C_{max}, and percentile, p, chosen. However, it will still be possible to add trains in step 2.

The train mix is given for the calculations in step 1, and thus the solution space will usually be quite small, due to the large train increments necessary to respect the train mix. A simple incremental approach is therefore used as the number of iterations are very few (in most cases below 5).

8.2.4 Step 2

Step 2: As described in step 1, the train mix induces large increments in the number of trains. There might thus be a large excess of capacity as the last feasible solution found in step 1 might have a infrastructure occupation, C, well below C_{max}. Thereby, this thesis assumes that adding a train of type i, will lead to a benefit of w_i. In step 2, the solution is therefore saturated from step 1 with additional trains. For this purpose the following mathematical model is proposed:

$$
\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{n}(w_i \cdot x_i) & \\
\text{subject to} \quad & x_i \geq 0 & \forall i \in \{1, 2, ..., n\} \\
& x_i + y_i = x_j + y_j & \forall \langle i, j \rangle \in Q \\
& C(X + Y, p) \leq C_{max} &
\end{aligned}
$$

Where $X = (x_1, ..., x_i, ..., x_n)$ is the decision variables for step 2. That is the number of extra trains of each train type. $Y = (y_1, ..., y_i, ..., y_n)$ is the number of trains found in step 1. Q is a set of tuples containing pairings between train types that represent the same service in opposite directions. For instance train type i offers a specific service between point A and B, while train type j offers the same service between B and A, thus

$x_i + y_i$ must be equal to $x_j + y_j$. This symmetry constraint may be neglected if symmetry is not required. w_i is the weight of each train type, with $1 \geq w_i \geq 0$ and $\sum_{i=1}^{n}(w_i) = 1$. The percentile, $p \in \{0, 1, ..., 100\}$, determines the ratio of possible sequences that should consume capacity less than or equal to the capacity threshold, C_{max}.

As described earlier, a high percentile will yield a lower capacity result as more train sequences has to be feasible compared to a low percentile. $C(X + Y, p)$ is the function (the model described in section 8.2.1 or an alternative) that calculates the infrastructure occupation given a set of trains, $X + Y$, and a percentile p.

The optimisation problem defined above looks simple. Unfortunately, it is not straightforward to solve since the constraint $C(X + Y, p) \leq C_{max}$ is computationally expensive to evaluate. Thus, to solve the problem an algorithm needs to be constructed that finds good solutions, and desirably the optimal solution, with a minimum amount of evaluations of $C(X + Y, p)$.

As described in section 6.2, a homogeneous train mix provides the possibility to run more trains than a heterogeneous train mix. Based on this characteristic, a greedy heuristic is used to search for solutions with as many trains of type as possible. The greedy heuristic is described in pseudocode in algorithm 2. The search works by considering the train types after weight in descending order. For each train type as many trains as possible are added. This will yield an upper bound on how many trains of a single train type can be added. This solution is subsequently improved, if possible, by adding as many trains as possible of the remaining types by weight. In a network such an improvement will generally be possible due to certain train types having little or no interaction with one another.

Algorithm 2: Pseudocode for greedy heuristic.

```
Data: Weights, w_i ∈ X
Result: X* = (x_1, ..., x_i, ..., x_n) a solution that maximize Σ_{i=1}^{n}(w_i · x_i)
x*_i ← 0   ∀i ∈ {1, ..., n}
for i ∈ X in descending order of w_i do
    x_i ← 0   ∀i ∈ {1, ..., n} // Reset solution
    x_i ← dSearch(i) // Find maximum amount of trains to add
    for j ∈ X in descending order by w_j do
        if j ≠ i then
            x_j ← dSearch(j)
        end
    end
    X* ← max(obj_val(X), obj_val(X*)) // Set best solution
end
return X*
```

To determine how many trains that can be added of single train type, given a start solution, a dichotomic search algorithm is used which is an extension of the well-known binary search algorithm. The search is a divide-and-conquer algorithm that works by dividing the search interval into two parts at each iteration.

For the problem of finding the maximum number of trains, the search works by initially determining a upper interval bound on the search interval. That is, a bound on the number of trains of a train type that yields an infeasible solution. Given this bound the best feasible solution is known to be in the interval between 0 and this upper bound. A good upper interval bound is close to the best solution. A bad upper interval bound will result in extra iterations, which are computationally expensive due to the sampling model used to evaluate the infrastructure occupation.

To test if it is even possible to add a train of the given type, the upper interval bound is initially set to 1. If one train yields an feasible solution, this is used as the lower interval bound, and the upper interval bound is increased by a guess. The guess is an estimate of how many trains that can be added of a single type. A good guess will provide a good upper interval search bound. Given that it is the same train type that is added, a guess could be based on the headway divided by the time period. This will result in the number of trains of the type in a total homogeneous situation.

When the upper interval bound has been found the search for the maximum number of trains is started. For each iteration the number of trains of the train type to be tested for feasibility is the midpoint between the lower interval bound and upper interval bound. If the tested number of trains yields an infeasible solution, the upper interval bound is set to the tested number of trains. If an feasible solution is found the lower interval bound is set to the tested number of trains. The algorithm is stopped after convergence and the best (maximum number of trains) is returned. This dichotomic search algorithm is also described in pseudocode in algorithm 3.

8.2.5 Solution space

The greedy algorithm used described in the previous sections finds the optimal solution if the solution space is concave and in certain cases convex. However, as the solution space might be neither convex nor concave, the optimal solution can be a point in the solution space that is not a corner point. To investigate how the solution space looks, two simple cases are considered where all feasible solutions are constructed. The two simple cases considered, are a line (only one edge) which is traversed by two heterogeneous train types and an extended line traversed by three train types. In the extended line which is

Algorithm 3: $dSearch(i)$ Pseudocode for dichotomic search.

```
Data: i ∈ X: train type for which as many trains as possible should be added
X = (x_1, ..., x_i, ..., x_n), Y = (y_1, ..., y_i, ..., y_n): current solution
p: percentile
C_max: infrastructure occupation threshold
Result: x_i^* maximum number of trains that can be added
x_i^* ← 0
guess ← guess on upper interval bound
intervalLowerBound ← 0
intervalUpperBound ← 1
/* Find a upper bound on how many trains of type i that can be added
   */
loop ← true
while loop do
    x_i ← intervalUpperBound
    // If infeasible
    if C(X + Y, p) > C_max then
        loop ← false // Upper bound for interval found
    else
        // If feasible
        intervalUpperBound = intervalUpperBound + guess
        if x_i > x_i^* then
            x_i^* ← x_i // Save as current best
        end
    end
end
/* Use binary search to find the maximum number of trains that can be
   added of type i                                                  */
x_i ← binarySearch(intervalLowerBound, intervalUpperBound)
return x_i^*
```

composed of three edges in sequence, one train type runs on all three edges and the two other runs only on a single, but different, edge. The line case is the simplest possible. The extended line case is an extension which is constructed to investigate the solution space when there are train types that does not have conflicts.

The simple line case is shown in figure 8.3. As observed in the figure the solution space is concave. This means that only the corner points compose the convex hull of the solutions given a linear objective function. Thus, the optimal solution will be either (12,0) or (0,12) depending on the weights defined and if no mix solution has been found in step 1. If a mix solution has been found in step 1, the solution space will be restricted (no train types with 0 trains) and the extreme points are therefore cut off (shown as a dashed

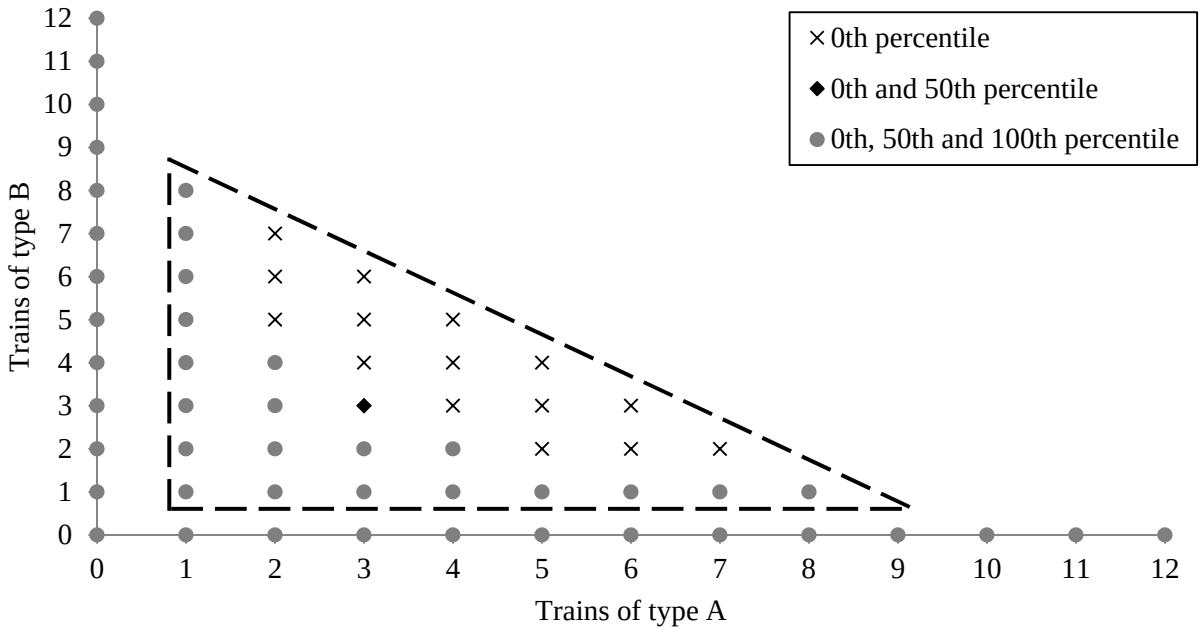


Figure 8.3: Solution space on a railway line with two heterogeneous train types. 0th, 50th and 100th percentile depicted. The dashed box depicts an arbitrary restricted solution space.

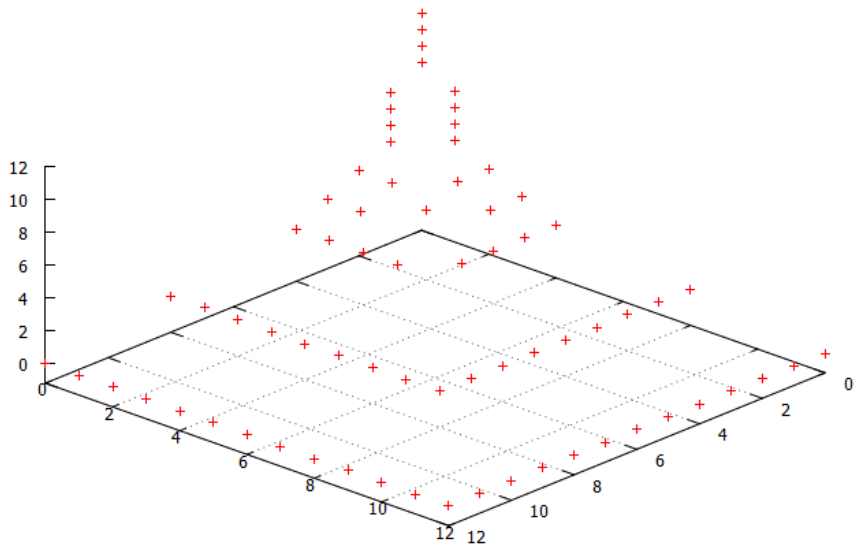


Figure 8.4: Solution space on an extended line with three heterogeneous train types. 50th percentile depicted.

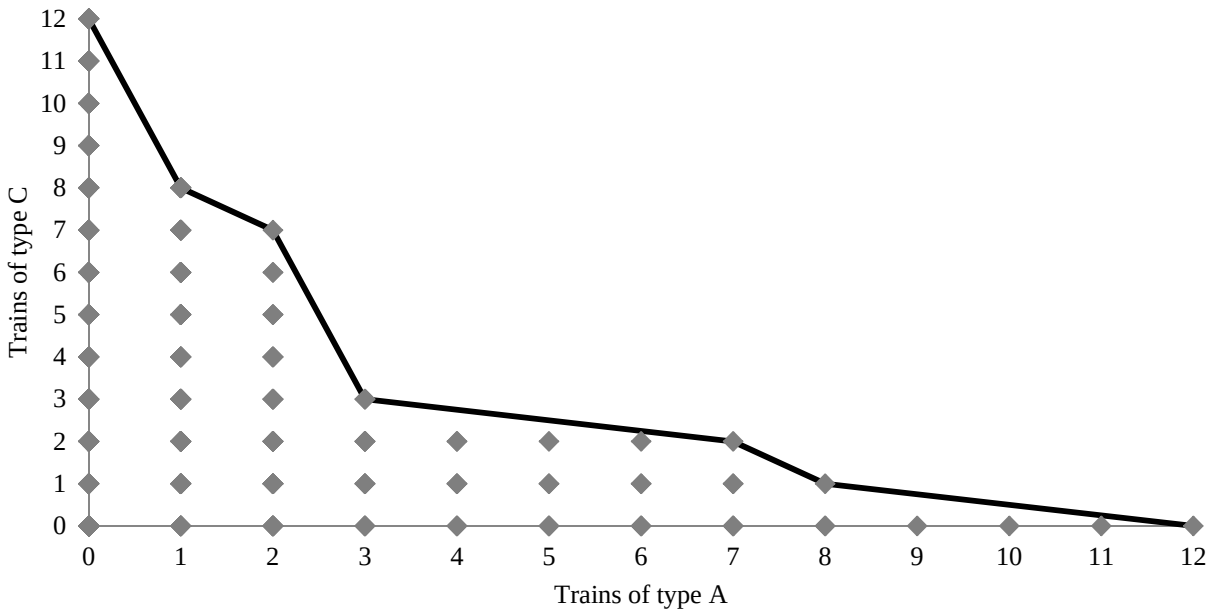


Figure 8.5: Solution space which is neither concave nor convex. 25th percentile for A,B-combination in the extended line case.

box in figure 8.3). However, the convex hull will still be composed by the corner points in the restricted solution space (due to the concave property).

The extended line case is shown in figure 8.4. As it can be observed in the figure the solution space is concave in two of three planes, just as for the line case. In the last plane the solution space is convex, which means that it is possible to add extra trains to some of the solutions in this plane without removing trains of the other types. This is caused by the fact that two of the train types only run on one edge which is not the same. It is therefore possible to saturate some of the solutions with extra trains. However, it should be noted that while this does not increase the infrastructure occupation, it will most likely increase the capacity consumption as buffer times are removed.

While figures 8.3 and 8.4 show solutions spaces which has desirable properties in form of a concave or rectangular (convex) solution space, then figure 8.5 shows a solution space that is neither convex nor concave. This solution space stems from the extended line case for the 25th percentile instead of the 50th percentile (median) depicted in figure 8.4. Thus in some cases the solution space will be neither convex nor concave and an optimal solution is therefore not necessarily composed by a corner point. Therefore, it can be concluded that the solutions found by the greedy heuristic are not necessarily optimal and thus optimality of the generated solutions cannot be guaranteed. To improve solutions, a meta-heuristic could be used, however, this will only provide the possibility for escaping local optimums (yielding better solutions), but will not ensure that a global

optimum is found. Alternatively, the whole problem can be represented by a mathematical model and solved using exact methods. However, as seen in previous approaches (section 8.1), this leads to a non-linear model, or a simpler model neglecting certain aspects (e.g. heterogeneity) captured by the approach suggested in this chapter.

8.3 Computational experiments

In this section, the proposed optimisation framework is applied to the network also used in chapter 7 (see figure 7.5 on page 116). In one hour, the goal is to estimate the number of trains the network can handle under a given mix plus additional trains maximising utility (sum of weights).

Table 8.1 shows an overview of the train types that should traverse the network and their route. The train types are heterogeneous in running time, especially between OD and SNO as described in section 7.4. For the case a minimum (block) headway time of 150 seconds is used. Exclusion constrains, as described in section 7.3.2.3, are not used.

Table 8.1: Train types in the network with route used, mix percentage for step 1 and weights for step 2 optimisation.

Train type	Route	Mix	Weight
1: Express train	1: OD(4) $\rightarrow$ FA(2)	1/16	3/26
2: $\hookrightarrow$	2: FA(6) $\rightarrow$ OD(2)	1/16	3/26
3: IC-A	1: OD(4) $\rightarrow$ FA(2)	1/16	2/26
4: $\hookrightarrow$	2: FA(6) $\rightarrow$ OD(2)	1/16	2/26
5: IC-B	3: OD(3) $\rightarrow$ ES(1)	1/16	2/26
6: $\hookrightarrow$	4: ES(3) $\rightarrow$ OD(1)	1/16	2/26
7: RE-A	5: OD(4) $\rightarrow$ FA(3)	1/16	1/26
8: $\hookrightarrow$	6: FA(5) $\rightarrow$ OD(1)	1/16	1/26
9: RE-B	7: FA(4) $\rightarrow$ ES(2)	2/16	1/26
10: $\hookrightarrow$	8: ES(3) $\rightarrow$ FA(1)	2/16	1/26
11: Freight	9: OD(3) $\rightarrow$ VM(2)	2/16	4/26
12: $\hookrightarrow$	10: VM(1) $\rightarrow$ OD(2)	2/16	4/26

This case network mainly serves as a basis for theoretical experiments to explore how the optimisation framework performs. In section 8.3.1 computational statistics are reported. In section 8.3.2, the mix and weights listed in table 7.2 is used to test different capacity thresholds and the appertaining capacity results. The optimisation is executed on a Windows 7 laptop with an Intel Core i7-2640M processor and 4 GB of RAM, and is implemented in C++.

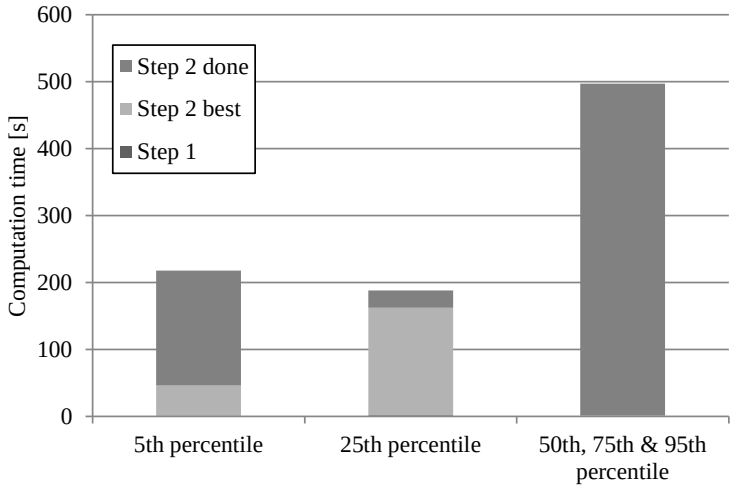


Figure 8.6: Computation time for each step in the algorithm.

8.3.1 Computational results

To investigate the computational performance of the optimisation framework the case is assesses with a threshold of 84% at the 5th, 25th, 50th, 75th and 95th percentile. As described earlier using the 0th and 100th percentile is not advisable due to variance in the capacity consumption model. Furthermore, the different percentiles represent how many percent of all permutations of train sequences that can be scheduled within the capacity threshold.

For the five percentiles calculated three unique results are found (see table 8.2). For the 5th and 25th percentile the given mix can be scheduled once according to the definition in table 8.1. In addition to this two extra trains of each types 3 and 4 can be added at the 5th percentile with an objective value of 0.21. For the 25th percentile, one additional train of each types 9 and 10 can be scheduled within the threshold with an objective value of 0.036. The latter is also the optimal solution as it is not possible to schedule additional trains of any other type in step 2.

For the 50th, 75th and 95th percentile, the mix cannot be scheduled. However, 18 trains of type 11 and 18 trains of type 12 can be scheduled with an objective value of 2.6 within a infrastructure occupation threshold of 84%.

Figure 8.6 shows the amount of time used in each step of the algorithm for the five percentile tested. In the figure, it can be observed that the calculation is completed significantly faster for the 5th and 25th percentile than for the 50th, 75th and 95th. The cause of the longer calculation time for the 50th, 75th, and 95th percentile compared to the 5th and 25th percentile is, however, not clear. The first step is completed equally fast

for all the five percentile, and the calculation terminates after 194-264 iterations for all the percentiles.

The best solution in step 2 is found significantly faster for the 50th, 75th and 95th percentile which is illustrated in figure 8.6 (1 second compared to 45 and 160 seconds). This is caused by the fact that for the 5th and 25th percentile the mix could be scheduled. This leads to a heterogeneous solution compared to the 50th, 75th and 95th percentile, where only two train types compose the solution. Due to this heterogeneity, the greedy heuristic is not able to find the best solution in the first couple of iterations as the train type with the highest weight cannot be added. This is, on the contrary, possible for the 50th, 75th, and 95th percentiles as the mix could not be scheduled in step 1 thus providing all the capacity for the step 2 optimisation.

8.3.2 Mix and capacity threshold

Using the developed optimisation framework, the capacity results for different infrastructure occupation thresholds for the 5th, 25th, 50th, 75th, and 95th percentile are investigated. The thresholds examined are 75%, 84% and 100%. The 75% is the UIC recommendation for lines with mixed traffic in the rush hour (UIC, 2013). 100% is the absolute maximum and can only be utilised under perfect conditions (maximum capacity), that is no delays can occur. The 75% by UIC is given for line sections and is not thus not an applicable threshold for routes and networks as the infrastructure occupation is higher when the whole network is considered instead of only a line section (see section 7.4.5). Therefore a threshold of 84% is also examined, which is based on an additional time rate of 18% derived in chapter 9 for the case network used in this chapter and chapter 7. This threshold value is considered to be the one reflecting practical network capacity best out of the three. Furthermore, mix and weights for the train types are differentiated as listed in table 8.1.

In table 8.2, the results of the capacity calculation is shown. Within a threshold of 75%, it is only possible to schedule the mix for the 5th percentile with an additional train of each of types 9 and 10. For the remaining percentiles at 75%, 16 trains of types 11 and 12 can be scheduled yielding an objective value of 2.3. These are the freight trains that have the highest weight of all the types. As the mix cannot be scheduled the final solution with the freight trains is very homogeneous and a total of 32 trains can therefore be scheduled.

For the 84% infrastructure occupation threshold it is possible to schedule the mix for the 5th and 25th percentile, but not the 50th, 75th and 95th percentile as also described

Table 8.2: Results with three different infrastructure occupation thresholds for five different percentiles. Numbers in parentheses is extra trains added in step 2. Objective values for step 2 marked in italic denote optimal solutions.

C_{max}	75%	75%	84%	84%	84%	100%	100%	100%	100%	100%	100%
Percentiles	95-25th	5th	95-50th	25th	5th	95th	75th	50th	25th	5th	
Mix scheduled		•		•	•		•	•	•	•	
Train type	Number of trains (extra trains added in step 2)										
1	0	1	0	1	1	0	1	1	1	1	
2	0	1	0	1	1	0	1	1	1	1	
3	0	1	0	1	3 (2)	0	1	1	1	1	
4	0	1	0	1	3 (2)	0	1	1	1	1	
5	0	1	0	1	1	0	1	1	1	1	
6	0	1	0	1	1	0	1	1	1	1	
7	0	1	0	1	1	0	1	1	1	1	
8	0	1	0	1	1	0	1	1	1	1	
9	0	3 (1)	0	3 (1)	2	0	3 (1)	2	2	2	
10	0	3 (1)	0	3 (1)	2	0	3 (1)	2	2	2	
11	16 (16)	2	18 (18)	2	2	22 (22)	2	3 (1)	4 (2)	5 (3)	
12	16 (16)	2	18 (18)	2	2	22 (22)	2	3 (1)	4 (2)	5 (3)	
Mix - objective	0 - 2.3	1 - *0.036*	0 - 2.6	1 -*0.036*	1 - 0.21	0 - 3.1	1 - *0.036*	1 - 0.14	1 - 0.29	1 - 0.43	

in the previous section. For the 100% threshold it is possible to schedule the mix for all percentiles except the 95th percentile. In addition to this it is possible to add extra trains in all solutions (usually freight trains as they have the highest weight) as listed in table 8.1.

For the 5th percentile at 75%, 25th percentile at 84%, and 75th percentile at 100% the optimal solution is found in step 2 as only one train of types 9 and 10 can be added.

Contrary to what could be expected, the total capacity increases for the 84% and 100% threshold with percentile. However, this is only when considering the number of trains. Results are as expected, when capacity figures are defined as the number of times a train mix can be scheduled plus additional trains that can be added given a certain objective value.

8.4 Summary and conclusions

In this chapter, an optimisation framework for the estimation of capacity in a railway network has been developed. The approach makes it possible to determine the number of times a certain mix of trains can be scheduled in a network within an infrastructure occupation threshold, C_{max}.

In addition to this mix, it is estimated how many trains that can be added according to predefined train type weights until the occupation threshold, C_{max}, is met. Differentiated weights make it possible to prioritise certain types as there may be higher demand for these.

As earlier described, in railway networks with heterogeneous operation, capacity consumption and infrastructure occupation values depend on the sequence (order) of trains. As the developed framework is able to take this into account it is a powerful tool for early (strategic) planning phases as it is able to give a span of capacity based on the relative amount of train sequences that should be feasible. Furthermore, the developed framework provides the means to ensure that the estimated capacity can yield an operation robust against delays if fully utilised (and timetabled properly).

For a Danish long distance network of 161 kilometres of double track lines, the capacity can be estimated in 3-5 minutes in most of the test instances used with optimisation framework. However, this may take up to 13 minutes in the worst case. Thus for strategic planning the model runs fast and should therefore be able to handle larger test instances than used for this chapter.

For future work, this thesis suggests three extensions that can be tested. Firstly, as the approach uses the infrastructure occupation as a robustness threshold, better estimates

on practical capacity could be obtained using values of capacity consumption instead of infrastructure occupation (as the simulation model of chapter 7 is thus used). Secondly, the second step problem formulation can be adjusted to distribute extra capacity among multiple train types. Lastly, meta-heuristics can be used instead of the greedy heuristic to find better solutions. Although, these extensions may lead to better solutions, they will also lead to an increase in computation times. For a more detailed discussion of the optimisations framework developed in this chapter, see chapter 10.

9 Relationships between capacity, heterogeneity, and robustness

In chapter 6 is was pointed out that capacity is not easily defined and assessed as it depends on how it is utilised. In section 6.2, four parameters were identified as significant for how capacity is utilised: the number of trains, heterogeneity, stability/robustness, and average speed.

The framework and model developed in chapter 7, provides the opportunity to assess different scenarios related to capacity and robustness. Specifically, this chapter seeks the answer to the following questions using the developed methodology:

1. What is the relationship between infrastructure occupation and capacity consumption (robustness) for networks – or maximum capacity versus practical capacity?

2. What is the relationship between (average) speed heterogeneity, infrastructure occupation, and robustness in networks?

The first question deals with the relationship between practical and maximum capacity; or in other terms the relationship between infrastructure occupation and capacity consumption. Here, the focus is to further investigate the network case described in section 7.4 and compare the results from this case with recommendations for line sections by UIC (2004b, 2013). As described in chapter 7, assessing networks (and routes) yields higher infrastructure occupation. This comes from the higher amount of dependencies in the network (and longer sections with (average) speed heterogeneity), but this also means that additional buffer times is implicitly added to the compressed network timetables. Thus this chapter seeks to study how this impacts the necessary additional time rate to be added to the infrastructure occupation to obtain the practical capacity (capacity consumption).

The second question deals with the buffer times that are made inherently available due to speed heterogeneity between trains. In very heterogeneous operation this leads to higher robustness compared to a more homogeneous operation. This may allow for a

higher infrastructure occupation as the practical capacity is often higher. As described in section 6.4.3 this is an aspect known in literature and practice for lines, although not studied in great detail. As described in the previous paragraph, analysing network means that more dependencies are taken into account thus providing even more "hidden" buffer time. For this second question this thesis thus seeks to find the relation between heterogeneity, infrastructure occupation, and robustness in networks and how this influence recommendations for infrastructure occupation, i.e. the relation between practical and maximum capacity.

The chapter is organised as follows. Section 9.1 describes the methodology and cases used, while the following sections 9.2 and 9.3 subsequently present and discuss the results obtained. Finally, conclusions and a summary are given in section 9.4.

9.1 Methodology

9.1.1 Models

The framework and model represented in chapter 7 is used as the tool to assess the relationships between different parameters of capacity as described in the introduction to this chapter.

In this regard the stochastic simulation model developed in this thesis (described in section 7.3.4) makes it possible to investigate the consecutive delays that generated train sequences suffer given a sampled input delay. This provides the possibility to assess the relationship between heterogeneity, infrastructure occupation, and consecutive delays (robustness) in networks (the second question listed in the introduction to this chapter). The output provided by the model is analysed visually using plots complemented by statistical measures to draw the conclusions.

For the theoretical analysis in section 9.3, no train sequences have been excluded when using the model of chapter 7 to make the results less biased for the regression analysis (i.e. the decisions of which sequences to exclude is a subjective decision). For the analysis in section 9.2, the same exclusions of undesirable sequences have been used as in the practical case of section 7.4.

9.1.2 Definition of measures

To investigate the impact of heterogeneity, an indicator of heterogeneity is needed. In section 4.2, a heterogeneity indicator (eq. 4.4) was put forward for headway heterogeneity in

cyclic timetables. Based on this indicator, this thesis proposes the following heterogeneity indicator for running time (average speed) heterogeneity in cyclic operation:

$$RTH = 1 - \left(\sum_{i=1}^{n-1} \min \left(\frac{t_i}{t_{i+1}}; \frac{t_{i+1}}{t_i} \right) + \min \left(\frac{t_1}{t_n}; \frac{t_n}{t_1} \right) \right) \cdot \frac{1}{n} \tag{9.1}$$

The heterogeneity indicator in eq. 9.1 can only be used on line sections. That is, on sections where the number of trains and the order of trains does not change. For a network consisting of a set of connected edges some aggregation therefore has to be done to obtain a single heterogeneity value for the network. Three simple aggregation methods in the form of a simple average, a weighted average, or a maximum as expressed in eq. 9.2, 9.3, and 9.4 are therefore tested in section 9.3.1, by this thesis, for their ability to capture running time heterogeneity in a network.

$$RTH^{Avg} = \frac{1}{|E|} \cdot \sum_{e \in E} RTH_e \tag{9.2}$$

$$RTH^{WAvg} = \frac{1}{|E|} \cdot \sum_{e \in E} n_e/n \cdot RTH_e \tag{9.3}$$

$$RTH^{Max} = max_{e \in E} RTH_e \tag{9.4}$$

9.1.3 Case networks

Two networks are used for the study conducted in this chapter. One is the network in Southwestern Denmark already described in section 7.4, where one of more of the four scenarios are used.

The other is a simple synthetic network (or corridor) as depicted in figure 9.1. As the figure shows, the network consists of four edges shaped like a Y. Five train types traverse the network with varying speed as represented by the running time in figure 9.1. Two trains of each type is added as an input to the model summing up to 10 trains total. The minimum headway time is set to 120 seconds. This simple network thus represents a corridor with converging routes and heterogeneous train operation, and is used for the analysis between heterogeneity and infrastructure occupation in section 9.3.1.

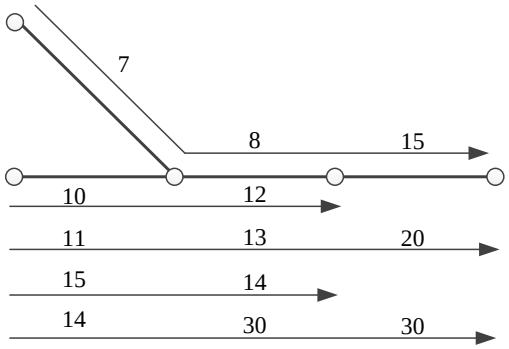


Figure 9.1: Synthetic network consisting of four edges in a Y-shape. The network is traversed by five different train types with route and running time (minutes) per edge depicted.

9.2 Practical and maximum capacity – the need for additional time rates

In section 6.1 it was described that the maximum capacity of a railway system cannot be utilised as this inevitably leads to delay propagation due to a lack of critical buffer times between trains. To account for this UIC (2004b, 2013) recommends that the infrastructure occupation rate should not exceed 60-85% depending on type of operation, see section 6.4.3 for more. These maximum infrastructure rates may be converted into additional time rates according to the formula:

$$\text{Additional time rate} = \frac{1}{\text{max. infrastructure occupation rate}} - 1 \tag{9.5}$$

UIC (2004b, 2013)'s recommendations for line sections as additional time rates are given in table 9.1.

Table 9.1: Additional time rates as suggested by UIC (2013) for line sections.

	Peak hour	Non-peak hour
Dedicated (sub)urban passenger traffic	18%	43%
Mixed traffic	33%	67%
Dedicated high-speed	33%	67%

In this section the additional time rates have been studied for the case network described in section 7.4, divided on four infrastructure scenarios. As set out in the introduction to this chapter, the recommendations given by UIC (2004b, 2013) is for line sections. Additional time rates for networks is expected to be lower due to "hidden" buffer times in

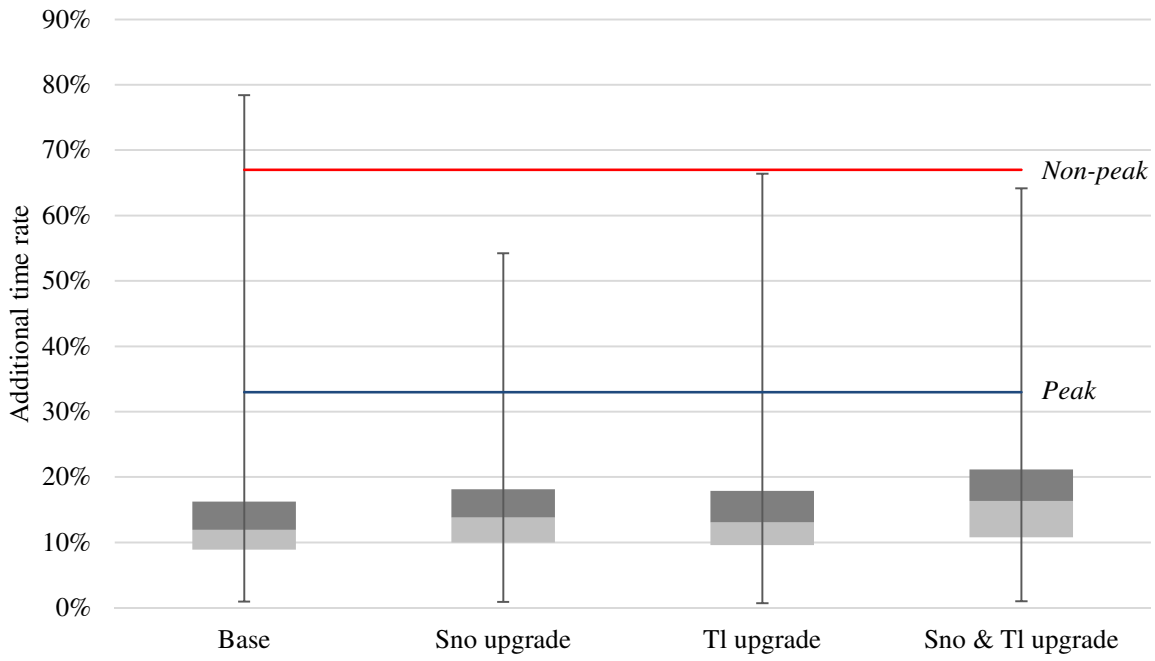


Figure 9.2: Additional time rates (maximum, 75th percentile, median, 25th percentile, and minimum, depending on critical buffer times needed) for the four infrastructure scenarios of the network in Southwestern Denmark (see section 7.4 for description). Horizontal lines depict (UIC, 2013) recommendations for mixed traffic.

the compressed network timetables (from which the infrastructure occupation is derived). This is caused by additional dependencies in a network compared to a decomposition into line sections.

Figure 9.2 shows box plots for additional time rates for the four infrastructure scenarios from the case study in section 7.4. The results have been derived by calculating the relative difference between infrastructure occupation and capacity consumption for each train sequence assessed by the model of chapter 7. Recall from section 6.4.2, that the difference between infrastructure occupation and capacity consumption is the addition of critical buffer times. The relative difference between infrastructure occupation and capacity consumption of a train sequence is thus the additional time rates that have to be added to the occupation time of a train sequence to account for robustness. This also defined in UIC (2013).

The results in figure 9.2 are illustrated as box plots as the additional time rates may differ according to the delay propagation generated by a sequence, and thus the critical buffer times needed, depending on the input delay sampled (and the sequence itself). The maximum value thus represents the additional time rate necessary to account for the maximum delay propagation across all train sequences, the 75th percentile for the necessary

additional time rate to account for the delay propagation of 75% of train sequences, and so on for the median, 25th percentile, and minimum.

Figure 9.2 shows three things. Firstly, it shows that the maximum values of additional time rates are much higher than the 75th percentile. This is not unexpected as it is caused by a few extreme cases. Therefore it is also recommended not to use the maximum additional time rate in operations planning as it will lead to an underutilisation of capacity. In this regard the 75th percentile serves as a better recommendation for the additional time rate.

Secondly, the figure shows that the 75th percentile is significantly lower than the UIC recommendations for both peak hour and daily (non-peak) operation. This result strengthens the hypothesis that ”hidden” buffer times in a network lead to a lower need of additional buffer times as described earlier in this chapter.

Lastly, it can be observed from the figure that the 75th percentile increase from the base scenario over the upgrade scenarios of Sno and Tl, respectively, to the upgrade of both Sno and Tl (from 16.2% to 21.2%). This is most likely caused by the fact that dependencies between trains are reduced (or removed) in the junctions. Thus some ”hidden” buffer times are removed resulting in a higher need of additional time rates (to add the necessary buffer times), again strengthening the hypothesis mentioned earlier on ”hidden” buffer times.

Given the results, a preliminary recommendation for an additional time rate for a network similar to the network of Southwestern Denmark could be the average of the 75th percentiles. That is 18.3%, which should be compared to the UIC recommendations for mixed traffic in the peak hour (33%).

9.3 Heterogeneity, infrastructure occupation, and robustness

A heterogeneous train operation induces implicitly added buffer times in the beginning or ending of line sections which yield additional buffer times and thus a reduced risk of delay propagation (as also mentioned in section 6.4.3 and the introduction to this chapter). On the other hand a high heterogeneity leads to higher infrastructure occupation as described in chapters 6 and 7.

In the following two sections, the relationship between heterogeneity, infrastructure occupation, and robustness is investigated for networks, not just line sections. In section 9.3.1, the correlation between heterogeneity and infrastructure occupation is studied. Furthermore, this section also studies the different aggregation methods for the hetero-

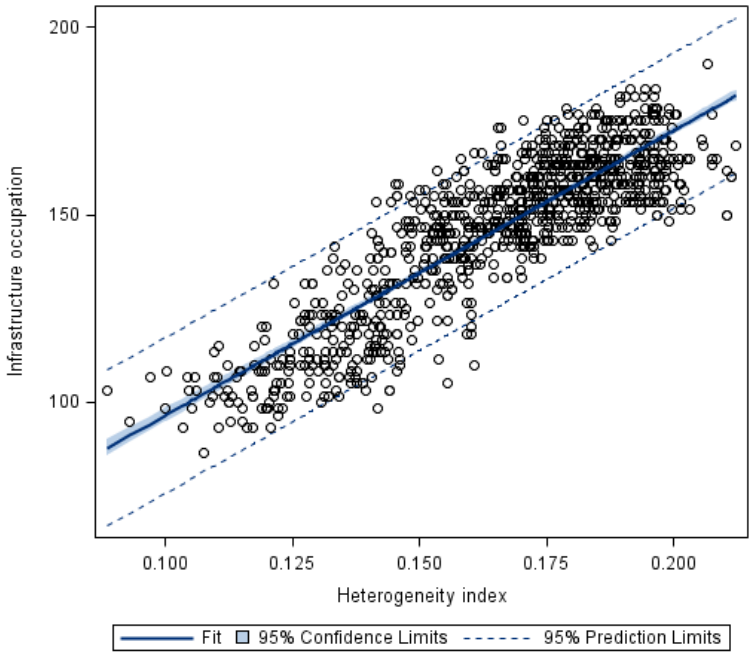


Figure 9.3: Linear regression for the simple network.

geneity indicators listed in section 9.1. Lastly, the relationship between heterogeneity and robustness is investigated in section 9.3.2 using the stochastic model output in the form of consecutive delays and the best heterogeneity indicator found in the analysis conducted in section 9.3.1.

9.3.1 Heterogeneity and infrastructure occupation in networks

As described earlier in this thesis there is a clear dependency between heterogeneity and infrastructure occupation on a line (section 6.2 and figure 6.5 on page 82).

To investigate the relationship between infrastructure occupation and running time heterogeneity in a network, a regression analysis has been conducted on the simple (Y-shaped) network described in section 9.1.3.

Specifically, the relation between infrastructure occupation and heterogeneity has been investigated as linear, (natural) logarithmic, reciprocal, and squared. For all three heterogeneity methods eq. 9.2, 9.3, and 9.4, this thesis finds that it is possible to obtain a good fit with both a linear, logarithmic, reciprocal, and squared regression model for the simple network. By examining residual plots, it is found that the linear and logarithmic model perform slightly better than the reciprocal model which in turn perform a little better than the squared model. Between the linear and logarithmic model, the difference is insignificant. As the linear model is the simplest, this model is suggested as a reflec-

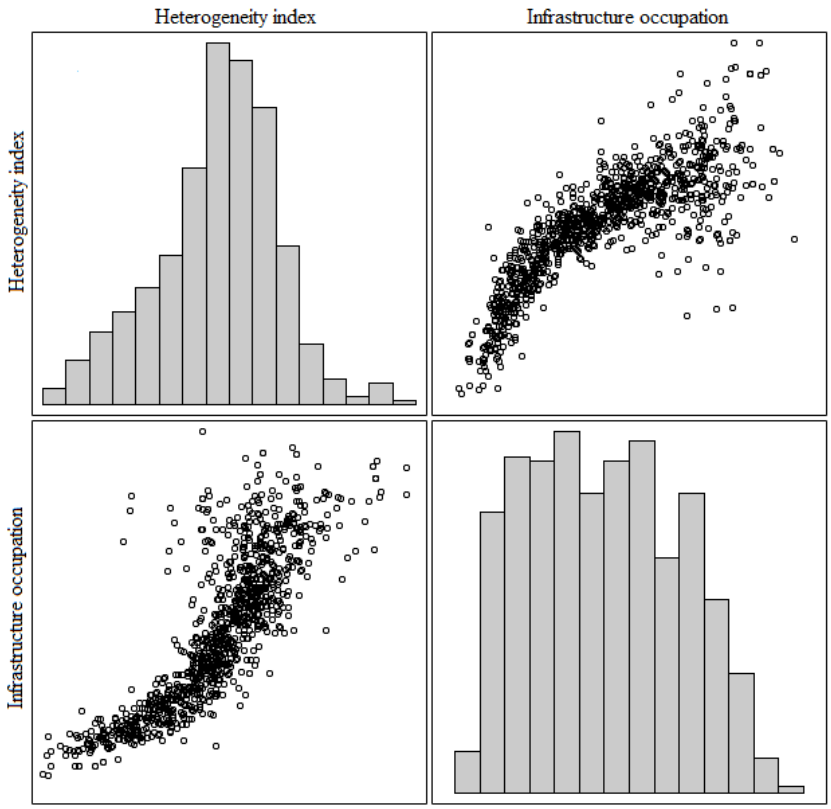


Figure 9.4: (Bottom left and top right) Correlation between infrastructure occupation and heterogeneity (index) for the network of Southwestern Denmark (Pearson correlation coefficient is 0.71). (Top left) Histogram showing the distribution of heterogeneity index values. (Bottom left) Histogram showing the distribution of infrastructure occupation values. Heterogeneity indices aggregated across edges using weighted average (eq. 9.3).

tion of the relationship between infrastructure occupation and heterogeneity in a simple network.

Fitted linear models for the three heterogeneity models result in a root mean square error (MSE) of 10.65, 10.62, and 12.02 for the average, weighted average, and the maximum aggregation methods. The R^2-values are, in the same order, 0.76, 0.76, and 0.69 for the three aggregation methods. Thus the weighted average, eq. 9.3, has the lowest MSE, and together with the average aggregation method the highest R^2 indicating that these heterogeneity indicators reflect the relationship between infrastructure occupation and heterogeneity better than the maximum aggregation. The difference between the average and the weighted average is small, and may be explained by the fact that the network is quite small and simple. Thus the extra information contained in the weighted average is not necessarily utilised. As the weighted indicator does contain extra information, this is used in the further analysis in this section.

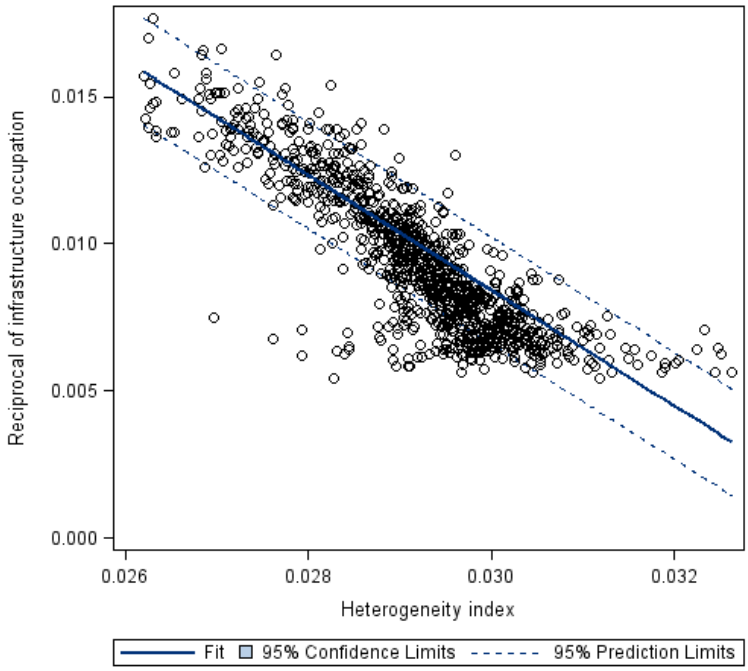


Figure 9.5: Reciprocal regression for the network of Southwestern Denmark (base scenario).

For the network in Southwestern Denmark, a regression analysis has been conducted similarly to the one for the simple network above. In figure 9.4 the correlation between infrastructure occupation and heterogeneity index using the weighted average is shown in the upper right and lower left part of the figure. The figure clearly shows that there is a non-linear relationship between infrastructure occupation and heterogeneity. Furthermore, the histogram for heterogeneity (upper left part of figure 9.4) shows that the majority of sequences have a high value of heterogeneity index, i.e. the median is larger than the mean. This also supported by the fact that the distribution is skewed to the right (skewness value of -0.33). The histogram for the infrastructure occupation in the lower right of figure 9.4 shows a more uniform distributions skewed towards lower values (skewness value of 0.59).

A regression analysis of the output for the Southwestern Denmark case base scenario shows that the best regression model is the reciprocal model followed by the natural logarithmic model. The linear and squared model does not provide a good fit. The fitted reciprocal model is shown in figure 9.5. As the figure shows, the reciprocal model does not model the non-linear relationship between infrastructure occupation and running time heterogeneity perfectly. The regression model may therefore be further improved by more advanced transformations (e.g. power transformations or splines). To maintain simplicity, the reciprocal relationship is accepted as a general model.

Lastly, it should be noted that assessing output data from the upgrade of both junctions instead of the base scenario leads to a more linear correlation between infrastructure occupation and heterogeneity. Again this is probably caused by the reduced amount of dependencies in the network. The increased linearity from the base scenario to the upgrade of both junctions is expressed by an increase in the Pearson correlation coefficient from 0.71 to 0.80 (where 1 is a perfect positive linear correlation, 0 is no correlation, and -1 is a perfect negative linear correlation).

9.3.2 Heterogeneity and robustness (consecutive delays)

Using the weighted average heterogeneity index described in eq. 9.3, the model of chapter 7, and the network of Southwestern Denmark, the relationship between heterogeneity, infrastructure occupation and consecutive delays have been calculated for a large sample of random sequences for the four infrastructure scenarios provided in the case.

The results of these calculations are summarised in figures 9.6 and 9.7, where figure 9.6 shows the results for the base network and figure 9.7 the results for the upgrade scenario of both junctions. The results depicted in the two figures represent the two extremes, why results for the individual upgrade scenarios of each of the junctions are not shown.

In the two figures, one point denotes the results (infrastructure occupation, heterogeneity index, and sum of consecutive delays) for a single sequence based on a single sample of input delays. That is one iteration out of the total amount of iterations completed. Thus the same sequence is present in the figures multiple times based on different samples of input delays. This leads to the vertical banding seen in both figures (especially in the right part of figure 9.6) as the only parameter changing is the consecutive delay with infrastructure occupation and heterogeneity index remaining constant (not affected by different samples of input delay).

From figure 9.6 two things can be observed. One is the large cluster to the left which seems to indicate that sequences with low heterogeneity leads to higher delay propagation than sequences with higher heterogeneity. The other is the cluster to the right which mainly consists of sequences with high heterogeneity. This cluster is probably caused by dependency in the network (a network effect), which is to some extent confirmed by figure 9.7 that shows a more linear relationship. The difference between the two being the lower amount of network dependencies due to the upgrade of both at-grade junctions.

In figure 9.8 the same data is represented as histograms (note that the consecutive delay is on a natural logarithmic scale). The figure shows the same tendency as figures 9.6 and 9.7, however more aggregated. From figure 9.8 it can be observed that high

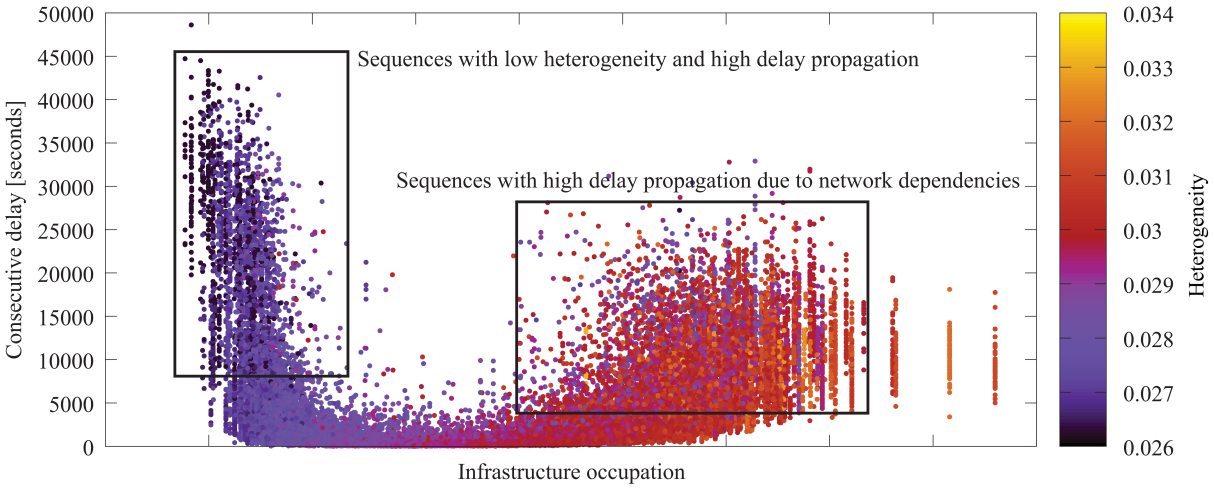


Figure 9.6: The relationship between infrastructure occupation (values increasing from left to right on the x-axis), heterogeneity index (as calculated using eq. 9.3), and consecutive delays for the base scenario in the Southwestern Denmark case. One point in the graph denotes one sequence based on a single sample of input delays (one iteration).

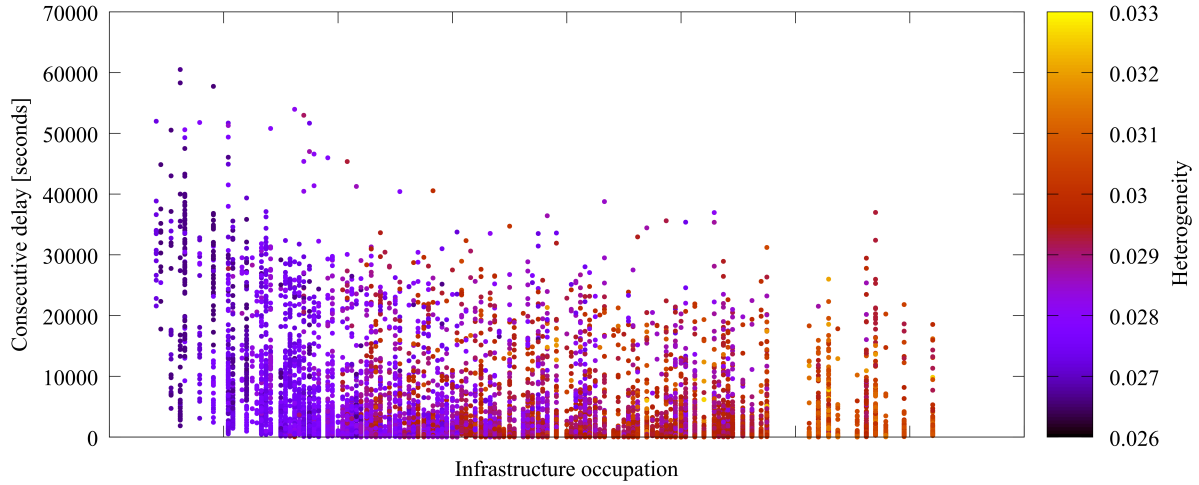


Figure 9.7: The relationship between infrastructure occupation (values increasing from left to right on the x-axis), heterogeneity index (as calculated using eq. 9.3), and consecutive delays for the upgrade scenario of both junctions in the Southwestern Denmark case. One point in the graph denotes one sequence based on a single sample of input delays (one iteration).

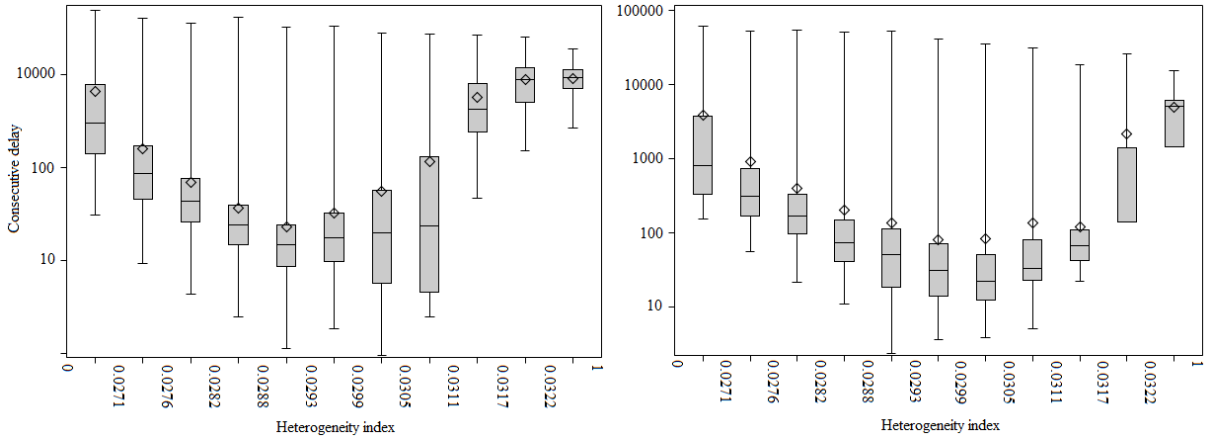


Figure 9.8: Histograms for base scenario (left) and scenario for the upgrade of both junctions (right). Box denotes first, second (median), and third quartile. End of whiskers denote minimum and maximum. Diamond symbols denote averages. Note that the y-axis is transformed by the natural logarithm.

heterogeneity (moving from left to right) leads to lower maximum delay propagation (top point on box plots), but not necessarily lower average delay propagation (denoted by diamonds). The tendency of higher heterogeneity leading to lower delay propagation is stronger in the upgrade case. This is ascribed to the reduced amount of dependencies in the upgrade case of Tl and Sno, as described several times earlier in this chapter.

The output on infrastructure occupation, heterogeneity, and consecutive delays has not made it possible to obtain a good regression model based on these parameters. However, correlation statistics for the output data show a negative Pearson correlation coefficient of -0.15 for the upgrade of both junctions and -0.09 for the base scenario. With a reciprocal transformation (similar to the one done in section 9.3.1) of the heterogeneity index the Pearson coefficients increase slightly to 0.11 for the base case and 0.16 for the upgrade of both junctions.

Acknowledging that there is some correlation between running time heterogeneity and robustness in networks, a link may drawn to the analysis of additional time rates in section 9.2 which can now be extended by the heterogeneity of train sequences.

An analysis of the results of such a relation indicates the same tendencies as seen earlier in this chapter. Thus, it is concluded that there is clearly a trend where higher heterogeneity leads to reduced maximum required additional time rates. That is the practical capacity is higher with a high heterogeneity if robustness against maximum delay propagation has to be handled. This does not imply that a train sequence with higher heterogeneity is always better than one with lower heterogeneity in networks, as this also depends on the dependencies in the network. However, if the only information available

is the heterogeneity index, it can be concluded that high heterogeneity leads to lower maximum delay propagation. Average delay propagation, contrary to maximum delay propagation, decrease with higher heterogeneity (negative correlation) to some point, after which the correlation turns positive as illustrated by the diamond symbols (the average) in figure 9.8. Again this is ascribed to network dependencies as the shift from negative to positive correlation occurs at higher heterogeneity index for the upgrade scenario. Thus on a line it is expected that the correlation will be negative across the whole range of heterogeneity indices.

9.4 Summary and conclusions

In the introduction to this chapter, the following questions were listed related to the relationship between robustness and capacity in networks:

1. What is the relationship between infrastructure occupation and capacity consumption (robustness) for networks – or maximum capacity versus practical capacity?
2. What is the relationship between (average) speed heterogeneity, infrastructure occupation, and robustness in networks?

These questions were studied by this chapter using the framework and model of chapter 7.

First, it was found that the additional time rates, which can be used to obtain the practical capacity (capacity consumption) from infrastructure occupation values, are lower in networks than the recommendations given for line sections by UIC (2004b, 2013). This is in line with expectations and is caused by the additional amount of buffer times available in the compressed network timetable. Complementing this, it is observed that the scenario with most dependencies (base) needs a lower additional time rate compared to the scenario with less dependencies (due to an upgrade of junctions). This is a consequence of less buffer times due to less dependencies.

For the case network of Southwestern Denmark, it was found that an additional time rate of approximately 18.3% is sufficient, where the corresponding recommendation from UIC (2004b, 2013) (line sections) is 33% for mixed traffic operation in the peak hour.

Secondly, for networks the relationship between heterogeneity and infrastructure occupation was investigated. First a heterogeneity indicator for running time heterogeneity was developed. Subsequently, it was found that for networks the relationship between the two parameters is best represented by a reciprocal function. For simpler network it is

found that a linear function also represents the relationship well. The correlation found between infrastructure occupation and heterogeneity is significant enough to be described by a (reciprocal or linear) model obtained using regression.

Using the developed heterogeneity index, the relationship between heterogeneity, infrastructure occupation, and robustness (consecutive delays) was investigated. It was found that train sequences with low heterogeneity has a risk of higher maximum delay propagation (consecutive delays) than train sequences with higher heterogeneity. The chapter also illustrates how network dependencies may lead to higher consecutive delays, for a large proportion of the possible train sequences.

End of part II

10 Discussion

This thesis contributes with the development of timetable complexities to indicate robustness of timetables at stations, a review of existing robustness indicators, and an evaluation of selected robustness indicators based on a comparison with microscopic simulations.

Furthermore, this thesis develops a framework to assess capacity consumed by a given set of trains in a network. This framework has additionally been extended with an optimisation framework to estimate the number of trains a given railway network can handle. Using the developed capacity consumption framework, a quantitative analysis was conducted on a case study in Southwestern Denmark to reveal capacity gains of infrastructure upgrades as well as relations between maximum and practical capacity in networks, and relationships between heterogeneity, robustness, and infrastructure occupation.

Chapters 4–5 and 7–9 describe these contributions in detail. The purpose of this chapter is to discuss these contributions, their practical uses, and suggestions for future work.

10.1 Robustness indicators

In chapter 4, a number of robustness indicators were identified, and a new indicator was developed to measure timetable complexity at stations. Robustness indicators have some clear advantages over simulation as they are computationally faster and usually requires less data input. In the early planning phases of strategic and tactical planning, this is important as it gives the possibility to assess alternative solutions faster and easier compared to simulations. Furthermore, it can be used in the objective functions of e.g. timetable optimisation instead of stochastic approaches.

A theoretical comparison of robustness indicators in this thesis showed (table 4.11 on page 61) that the reviewed and developed indicators are not able to capture all parameters that affect robustness individually (as a simulation is able to). That is the size and distribution of time supplements, buffer times, and initial delays. However, what the

study did show was that more robustness indicators combined can theoretically capture all aspects of robustness.

An aggregated indicator could therefore be interesting to investigate in the further work and could possibly be obtained using regression on a number of different cases. The aggregated indicator could for instance be obtained by simply using timetable complexity and risk profiles as predictors. Timetable complexities then captures the risk of delay propagation at stations based on the size and distribution of buffer times and delays, and risk profiles represent the amount of time supplements available between stations to absorb delay. Such an aggregated index will probably perform in a satisfying way for lines. For networks, the picture will be more blurry as network complexity (effects of route and train dependencies) has a large impact on results (as e.g. shown in section 9.3). A way to solve this could be a network measure to indicate the effect the characteristics/complexity of a network has on robustness.

In chapter 5, different robustness indicators were evaluated to uncover their ability to indicate robustness and the semantics of such indicators. It was found, that the indicators correlate well with the simulation results except the headway heterogeneity indices. This was mainly ascribed to the fact that the scenarios studied did, mainly, contain a lot of running time supplement. This probably led to buffer times not being used where the headway heterogeneity indices is an indicator the distribution of the buffer times. However, actual semantics of the indicators could not be obtained. To uncover the actual semantics of the indicators better, significantly more infrastructure and timetable scenarios are needed to obtain a larger data set. This could make it possible to recommend values of indicators to IMs and RUs for timetable and infrastructure planning purposes. Furthermore, it could be implemented in optimisation models to stop optimisation when a given robustness criteria have been met and other parameters can then subsequently be optimised.

10.2 The capacity models and their applications

The framework developed in chapter 7 for strategic assessment of infrastructure occupation and capacity consumption in railway networks may be used with any method(s) found suitable. Even a manual approach can be used, although this would be cumbersome to use in practice. This thesis, on the other hand, developed an efficient model implementing the framework.

As the model generates and assesses different train sequences based on a set of trains, taking heterogeneity into account, it is a powerful tool for strategic capacity planning

as a predefined timetable is not necessary while service intentions and their implications (heterogeneity) on capacity is still modelled, contrary to many existing approaches. For strategic planning, possible scenarios to be assessed will usually be infrastructure scenarios, as done in chapter 7, but can also include assessment of capacity ramifications when using new or alternative rolling stock, or new line plans/service intentions.

For future work on this model, this thesis suggests to improve the model by scanning the generated sequences for possible overtakings that will reduce infrastructure occupation (and capacity consumption). Overtakings can be handled by the present model, but this does not work well in practice due to undesirable results (infrastructure occupation may increase instead of decrease).

Furthermore, as the model runs fast, the next step is naturally to test larger networks and especially to investigate how to handle exclusions of the many undesirable sequences in such networks. In this context, an approach where desirable train sequences are found using a search, rather than the Monte-Carlo sampling done, could offer potential in terms of computational speed and success rate in finding desirable sequences. However, as mentioned in section 7.3.2.3, a search implies that sequences obtained are not random as opposed to the sampling approach, thus producing a biased distribution of capacity consumption. This either has to be accepted, or an unbiased search procedure has to be devised.

The capacity determination (optimisation) framework developed in chapter 8 provides estimates on the number and type of trains a given network can handle. As the optimisation framework is based on the capacity consumption model in chapter 7 it can handle heterogeneity of train types contrary to approaches based on a single or average train type. The developed optimisation framework may be used to quantify capacity gains in terms of number of trains instead of capacity consumption (or infrastructure occupation).

Robustness is handled in the optimisation framework in the form of a infrastructure occupation threshold to ensure the proper addition of buffer times. Thus the practical capacity is assessed and not the maximum capacity, although it is possible to assess the maximum capacity by setting the threshold to 100%. The stochastic simulation of chapter 7 is not used in the optimisation framework. Thus better estimates on practical capacity could be obtained using capacity consumption values rather than infrastructure occupation values, i.e. using the stochastic simulation to estimate critical buffer times.

However, in this context computational speed can be an issue as obtaining infrastructure values is significantly faster than obtaining capacity consumption values (e.g. approx. 10 seconds vs. 15 minutes as described in sections 7.4.5 and 7.4.6). A solution may be

to use the stochastic simulation only in some iterations of the optimisation, to produce infrastructure occupation thresholds to be used in the other iterations.

For the second step of the capacity determination, a linear objective function is used to maximise the number of trains based on differentiated train type weights. The linear objective is not desirable in all cases as trains types with the highest weights are always sought to be added first.

To find the best solutions in this second step a greedy heuristic was used, which given the objective function used is a suitable approach. However, this heuristic does not necessarily find the best solutions as the solution space is not concave nor convex in all cases. Thus leading to the possibility that only local optimums are being found.

For future work, this thesis suggests to adjust the second step problem formulation, so train type weights decrease logarithmically by the number of trains added, to ensure that not only one train type gets added. To solve this revised problem, the greedy heuristic is not ideal (in addition to the fact that it cannot escape local optimums). This thesis therefore proposes to use meta-heuristics with the alternative problem formulation. Although, this may lead to an increase in computation times.

In chapter 9, the relationship between infrastructure occupation and capacity consumption (or maximum and practical capacity) for networks was investigated and compared to UIC 406 line divisions using the capacity consumption model from chapter 7. It was found that additional time rates used to obtain the practical capacity is lower for a network than for line sections. Furthermore, the relationship between heterogeneity, infrastructure occupation and robustness was studied, where is was found that dependencies in a network seem to cause a non-linear relationship between infrastructure occupation and heterogeneity, contrary to a linear relationship on line sections and networks with a low amount of dependencies.

For these studies, one case was used. For future work, additional studies of a large variety of cases (networks) could provide precise estimates on practical versus theoretical capacity on line sections, routes, and networks. Furthermore, the non-linear relationship between capacity parameters in networks (network effect) could possibly be expressed by one of the complexity indices described in section 4.6. Alternatively, a new indicator specifically describing the complexity (dependencies) of a network could also be developed.

In addition to the case network used in this thesis, the capacity consumption framework has also been used in an analysis of a network between Copenhagen and Nykøbing F in Denmark as described in Jensen and Rasmussen (2015). The results are shown in figure 10.1. Without going into details, the results depicted show two things: infrastructure upgrades increase the relative number of feasible train sequences (out of all possible

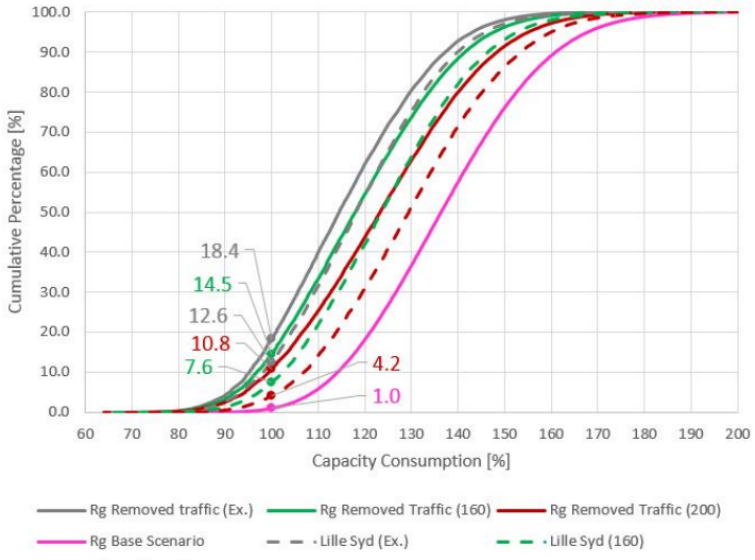


Figure 10.1: Results of a capacity analysis for the corridor between Copenhagen (Ny Ellebjerg) and Nykøbing F using the framework described in chapter 7 (Jensen and Rasmussen, 2015). Values indicate the percentage of feasible (maximum 100% consumption) train sequences out of all possible train sequences.

train sequences), and higher speeds (200 km/h versus 160 km/h) result in a lower relative amount of feasible train sequences caused by an increase in speed heterogeneity between passenger and freight trains. Thus the capacity consumption model developed by this thesis proved able to capture the benefit of infrastructure upgrades and the heterogeneity of train types.

Finally, the models developed may be used to further investigate the relationship between operational parameters for capacity, i.e. heterogeneity, number of trains, average speed, stability/robustness and infrastructure occupation, as to some extent started in chapter 9, where the stochastic simulation was used to conclude that sequences with a low heterogeneity has an increased risk of high delay propagation.

The works of Krueger (1999); Abril et al. (2008); Landex (2008a); and Harrod (2009) have also dealt with this to some extent. However, the main focus of these authors is on lines or simple networks/corridors. Thus there is research gap that the models developed in this thesis may close if used on a set of different networks and operational parameters. Such a study could also result in a function (based on regression) that potentially could improve computation speeds of optimisation models, like the capacity determination framework developed in chapter 8. This is also an approach used in simulation optimisation to reduce the number of times a simulation-based objective function is evaluated. Furthermore, it could provide recommended values for planners.

10.3 Mesoscopic simulation model

In section 7.3.4, this thesis developed an event-driven stochastic simulation model based on a mesoscopic infrastructure model. The infrastructure model used in this simulation model includes station routes and thus the simulation is able to model delay propagation occurring in station switch zones due to routes conflicts. As described in chapter 7, the model is fast as it runs in linear time in the order of events (departures/arrival of each train at vertices), while efficiently taking advantage of multi-threading in multi-core processors. For instance simulating one compressed timetable in the network used in section 7.4 takes less than 0.01 second per processor thread for 60 iterations on a laptop. Using a multi-core processor this number is even lower as iterations are divided among the cores/threads.

The speed of the simulation makes other applications of the model interesting. On a network level, the model is particularly interesting compared to the robustness indicators also dealt with in this thesis. As stated earlier, the robustness indicators may be used with some effort in networks, however, a simulation model will provide more precise results for networks as no decomposition is made. On the other hand, the simulation model may also require more input data.

Given the fast computation time, the simulation model may be used for timetable optimisation. Possibly, only in some iterations with the computationally faster robustness indicators providing robustness estimates in the other iterations. For stations, the infrastructure/timetable complexities described and developed in this thesis will typically provide better estimates than the simulation model, as more routes/details of the switch zone(s) are included.

There is thus a large potential to harvest by integrating indicators and simulation to obtain even faster and more precise models, although this will require a considerable effort to implement, calibrate, and test.

Lastly, the simulation model may be integrated into other models. For instance, simulation of train delays in traffic models to yield better estimates on robustness benefits when upgrading railway infrastructure.

11 Conclusions

With a railway industry continuously striving to reduce cost and utilise resources optimally, as well as a focus on achieving faster travel times, higher frequency, and better robustness against delay, there is an ongoing need to advance current decision support systems.

With the focus on the strategic and early tactical planning phases, the purpose of this PhD study has therefore been to review, evaluate, and develop indicators that can indicate the robustness of railway operation and infrastructure plans. Furthermore, it has been the purpose to develop a methodology for assessment of (robust) capacity in railway networks independent of a predefined timetable with only the service intentions (or plan of operation) given.

The initial focus in this thesis was on robustness indicators, which was chosen as a subject as they can be used as a computationally fast alternative to simulation, usually with low input requirements. In the thesis, the following robustness indicators are initially identified: headway heterogeneity indices, UIC 406 compression results, distribution of running time supplements, network locations with a high risk of delay propagation, and timetable and infrastructure complexities. However, a theoretical comparison reveals that none of the reviewed indicators are able to capture initial delays, buffer times, and time supplements, which are parameters of robustness, at the same time. However, the various indicators reviewed are able to capture some of these parameters. For instance, the UIC 406 method can indirectly be used to indicate the amount of buffer times available. While headway heterogeneity measures can be used as an indicator for distribution of buffer times.

Based on the review, a timetable complexity index was developed by this thesis that contributes by extending existing infrastructure complexities. The timetable complexity index can be used to assess robustness of a timetable at stations based on the infrastructure layout, timetable, and a probability distribution for delays. In a case study

on the Danish railway station of Skanderborg, it was confirmed by this thesis that the timetable complexity developed is more accurate and detailed than previous approaches when comparing different timetable alternatives.

Subsequently, this thesis evaluated the following robustness indicators that represent the whole range of robustness parameters: headway heterogeneity, infrastructure and timetable complexity indices, UIC 406, train path risk profiles and compared these with a microscopic simulation for their ability to indicate robustness. Based on the comparison this thesis concludes that the majority of the evaluated indicators correlate well at indicating robustness at a relative scale. At an absolute scale recommendations are harder to make for obtaining a given punctuality goal. Headway heterogeneity measures did not correlate well with simulation results which can be ascribed to the nature of the case (high amount of running time supplements, as captured by the risk profile indicator).

The main contribution of this thesis is a framework to evaluate capacity consumption in railway networks in a way not previously done. Contrary to existing approaches the framework is able to handle networks and robustness given only the infrastructure and a plan of operation in the form of number and type of trains (with routes, headways and running times). This reduces the work-load in the strategic planning phase, thus enabling a planner to evaluate multiple scenarios fast.

As the whole network is evaluated, with the proposed framework, the deficiencies of existing methods for capacity assessment is dealt with. For instance, the UIC 406 method uses decomposition into line sections which may lead to ambiguous results.

The capacity consumption framework was implemented with a model using an adjusted longest path search to find the infrastructure occupation in a headway graph (representing the required headway times between trains). Furthermore, the model uses a mesoscopic infrastructure model to ensure fast computation times while allowing to use input data of varying magnitude and precision. The model generates a number of compressed timetables for a given network and set of trains from where the infrastructure occupation is derived using the longest path found. Using a stochastic discrete event simulation, the compressed timetables are subsequently simulated to produce a distribution of capacity consumption.

The developed model was tested on a number of infrastructure scenarios on a network in Southwestern Denmark. These tests showed that the model can capture the capacity and robustness gain in different infrastructure scenarios.

Additionally, this thesis developed an optimisation framework to determine the number of trains that can enter a given network under robust capacity utilisation. The devel-

oped optimisation framework takes the heterogeneity of operation into account, contrary to many existing approaches, and is thus a powerful tool for early (strategic) planning phases as it is able to give a span of capacity. The framework is a bi-level (simulation-based) optimisation approach that utilises the capacity framework described above. The optimisation framework consists of two steps. First, trains are added to the network according to a predefined mix of train types. Secondly, the solution from the first step is saturated with extra trains using a heuristic. The optimisation framework was tested on the network in Southwestern Denmark, which showed that the capacity could be estimated in 3-5 minutes in most test instances.

In this thesis, infrastructure occupation and capacity consumption values for networks was, furthermore, compared with the line division values known from the UIC 406 method using the developed capacity consumption model. Additionally, the relation between heterogeneity, robustness, and infrastructure occupation was also investigated.

This thesis finds that the additional time rates, which can be used to obtain the practical capacity (capacity consumption) from infrastructure occupation values, are lower in the tested network of Southwestern Denmark than the recommendations given for line sections by the UIC (18.3% compared to 33%). This is caused by additional buffer times inherently available in the compressed network timetable compared to compressed timetables on line sections. This is also expressed through the correlation between train dependencies in the network and additional time rates required in different scenarios assessed by this thesis.

Subsequently, this thesis developed a running time heterogeneity index for networks based on a weighted average and found that for networks the relationship between heterogeneity (index) and infrastructure occupation is best represented by a reciprocal function. Although, it was also found that a linear function represents this relationship well for a simpler network.

Furthermore, this thesis found that train sequences with a low heterogeneity have an increased risk of higher maximum delay propagation (consecutive delays) compared to train sequences with higher heterogeneity. This is caused by additional buffer times that are implicitly added to heterogeneous sequences, thus increasing their robustness against delay propagation, where more homogeneous sequences have less additional buffer times. Lastly, this thesis also found that network dependencies may lead to higher consecutive delays for a large proportion of the possible train sequences.

For future work, applying robustness indicators to more networks and scenarios to uncover semantics of the indicators using simulation is of interest. In this context, it is especially of interest to investigate if an aggregated robustness indicator can be derived trough regression on existing, and possibly new, indicators used on the results of additional cases.

Regarding the capacity consumptions model, this thesis suggests to improve the model by scanning the generated sequences for possible overtakings that will reduce infrastructure occupation and capacity consumption.

Furthermore, as the model runs fast, the next step is naturally to study larger networks and how to handle combinatorial problems related to the exclusion of undesirable train sequences in such networks. In this context, an unbiased search to find desirable train sequences will be preferably over the Monte-Carlo sampling done. However, deriving an unbiased search is not straight-forward.

For future work on the capacity determination (optimisation) framework, this thesis suggests to test the use of the stochastic model of chapter 7 for more precise robustness estimates, and to adjust the second step problem formulation so train type weights decrease logarithmically with the number of trains added. The latter will ensure that not only one train type gets added, unless desired. To solve this revised problem, the used greedy heuristic is not ideal and this thesis therefore proposes to use meta-heuristics instead.

Additionally, the capacity models developed also lay the foundation for further research as it provides not only a methodology useful in practice, but also the means for researching capacity parameters and robustness of operation easily (as briefly done in chapter 9). For instance, extensive research can be conducted into the relationships between heterogeneity, number of trains, average speed, stability/robustness and infrastructure occupation in networks.

Lastly, the fast computational speed of the simulation model developed makes other applications of the model useful. For instance for robustness assessment in timetable optimisation and traffic models on a network level.

References

Assessing travel time reliability in transport appraisal. *Journal of Transport Geography*, 18(3):419 – 425, 2010. ISSN 0966-6923.

M. Abril, F. Barber, L. Ingolotti, M. Salido, P. Tormos, and A. Lova. An assessment of railway capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(5):774–806, 2008.

T. Albrecht. *Energy-Efficient Train Operation*, chapter 5, pages 83–105. Eurailpress, Hamburg, Germany, 1st edition, 2008.

E. V. Andersson, A. Peterson, and J. Törnquist Krasemann. Quantifying railway timetable robustness in critical points. *Journal of Rail Transport Planning & Management*, 3(3):95–110, aug 2013. ISSN 22109706. doi: 10.1016/j.jrtpm.2013.12.002.

E. V. Andersson, A. Peterson, and J. Törnquist Krasemann. Reduced railway traffic delays using a MILP approach to increase Robustness in Critical Points. *Journal of Rail Transport Planning & Management*, 5(3):110–127, nov 2015. ISSN 22109706. doi: 10.1016/j.jrtpm.2015.09.004.

J. Armstrong, J. Preston, and C. Potts. Developing Capacity Utilisation Measures and Limits for Railway Nodes. In *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis*, Copenhagen, 2013. IAROR.

Banedanmark. SODB anlægsbestemmelser. Technical report, Copenhagen, 2013.

J. P. Bendfeldt, U. Mohr, and L. Müller. RailSys, a system to plan future railway needs. In *Computers in railways VII*, pages 249–255, 2000. ISBN 1-85312-826-0.

N. Besinovic, R. Roberti, E. Quaglietta, V. Cacchiani, P. Toth, and R. M. P. Goverde. Micro-macro approach to robust timetabling. In *Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis (IAROR): RailTokyo2015*, Tokyo, mar 2015.

R. A. Brualdi. *Introductory Combinatorics*. Elsevier North-Holland, inc., 1977. ISBN 0-7204-8610-6.

O. Brünger and E. Dahlhaus. *Running Time Estimation*, chapter 4, pages 58–82. Eurailpress, Hamburg, Germany, 1st edition, 2008.

T. Büker and B. Seybold. Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning & Management*, 2(1-2):34–50, 2012. ISSN 22109706. doi: 10.1016/j.jrtpm.2012.10.001.

R. Burdett and E. Kozan. Techniques for absolute capacity determination in railways. *Transportation Research Part B: Methodological*, 40(8):616 – 632, 2006. doi: 10.1016/j.trb.2005.09.004.

V. Cacchiani and P. Toth. Nominal and robust train timetabling problems. *European Journal of Operational Research*, 219(3):727–737, jun 2012. ISSN 03772217. doi: 10.1016/j.ejor.2011.11.003.

G. C. Caimi. *Algorithmic decision support for train scheduling in a large and highly utilised railway network*. PhD thesis, ETH Zürich, 2009.

M. Carey and S. Carville. Scheduling and platforming trains at busy complex stations. *Transportation Research Part A: Policy and Practice*, 37(3):195–224, Mar. 2003. ISSN 09658564. doi: 10.1016/S0965-8564(02)00012-5.

A. Ceder. *Public transit planning and operation: theory, modeling and practice*. Elsevier, Butterworth-Heinemann, 1st edition, 2007. ISBN 978-0750661669.

F. Corman, A. D'Ariano, M. Pranzo, and I. Hansen. Effectiveness of dynamic reordering and rerouting of trains in a complicated and densely occupied station area. *Transportation Planning and Technology*, 34(4):341–362, 2011. doi: 10.1080/03081060.2011.577152.

Danish Ministry of Transport. En grøn transportpolitik (29-01-2009), 2009. URL http://www.trm.dk/da/politiske-aftaler/2009/en-groen-transportpolitik.

A. de Kort, B. Heidergott, and H. Ayhan. A probabilistic (max, +) approach for determining railway infrastructure capacity. *European Journal of Operational Research*, 148 (3):644 – 661, 2003. ISSN 0377-2217. doi: 10.1016/S0377-2217(02)00467-8.

T. Dewilde. *Improving the Robustness of a Railway System in Large and Complex Station Areas*. PhD thesis, KU Leuven, dec 2014.

T. Dewilde, P. Sels, D. Cattrysse, and P. Vansteenwegen. Improving the robustness in railway station areas. *European Journal of Operational Research*, 235(1):276–286, may 2014. ISSN 0377-2217. doi: 10.1016/j.ejor.2013.10.062.

European Commission. White paper: Roadmap to a single european transport area– towards a competitive and resource efficient transport system, 2011.

European Commission. First railway package of 2001, 2012. URL http://ec.europa.eu/transport/rail/packages/2001{_}en.htm.

M. Fischetti, D. Salvagnin, and A. Zanette. Fast Approaches to Improve the Robustness of a Railway Timetable. *Transportation Science*, 43(3):321–335, aug 2009. ISSN 0041-1655. doi: 10.1287/trsc.1090.0264.

A. Gille, M. Klemenz, and T. Siefer. Applying multiscaling analysis to detect capacity resources in railway networks. *Timetable Planning and Information Quality*, page 73, 2010.

R. M. Goverde. Timetable Stability Analysis. In I. A. Hansen and J. Pachl, editors, *Railway Timetable & Traffic*, chapter 7, pages 118–134. Eurailpress, Hamburg, Germany, 1st edition, 2008.

R. M. Goverde and I. A. Hansen. Performance indicators for railway timetables. In *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings*, pages 301–306. IEEE, aug 2013. ISBN 978-1-4673-5277-2. doi: 10.1109/ICIRT.2013.6696312.

R. M. Goverde, I. A. Hansen, G. Hooghiemstra, and H. P. Lopuhaa. Delay Distributions in Railway Stations. In *The 9th World Conference on Transport Research*, number 3605, 2001.

R. M. Goverde, F. Corman, and A. D'Ariano. Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions. *Journal of Rail Transport Planning & Management*, 3(3):78–94, Aug. 2013. ISSN 22109706. doi: 10.1016/j.jrtpm.2013.12.001.

R. M. P. Goverde. *Punctuality of railway operations and timetable stability analysis*. PhD thesis, Delft, the Netherlands, 2005.

R. M. P. Goverde. Railway timetable stability analysis using max-plus system theory. *Transportation Research Part B: Methodological*, 41:179–201, 2007. ISSN 01912615. doi: 10.1016/j.trb.2006.02.003.

O. Happel. Sperrzeiten als grundlage für die fahrplankonstruktion. *Eisenbahntechnische Rundschau*, 8(2):79–90, 1959.

N. Harris. Train boarding and alighting rates at high passenger loads. *Journal of Advanced Transportation*, 40(3):249–263, 2006.

S. Harrod. Capacity factors of a mixed speed railway network. *Transportation Research Part E: Logistics and Transportation Review*, 45(5):830–841, sep 2009. ISSN 13665545. doi: 10.1016/j.tre.2009.03.004.

L. Herrstedt. Dimensionsgivende Trafikant Reaktionstid. Technical report, Trafitec, Lyngby, 2007. URL http://trafitec.dk/sites/default/files/publications/reaktionstid.pdf.

M. Hofmann. On the Complexity of Parameter Calibration in Simulation Models. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 2(4):217–226, oct 2005. ISSN 1548-5129. doi: 10.1177/154851290500200405.

D. Huerlimann, G. Longo, and G. Medeossi. Stochastic micro-simulation as a timetable robustness estimation tool. In *RailZürich*, Zürich, 2009.

T. Huisman, R. Boucherie, and N. van Dijk. A solvable queueing network model for railway networks and its validation and applications for the netherlands. *European Journal of Operational Research*, 142(1):30 – 51, 2002. doi: 10.1016/S0377-2217(01)00269-7.

J. Jacobs. *Rescheduling*, volume 1, chapter 11, pages 182–191. Railway Timetable & Traffic, Hamburg, Germany, 1st edition, 2008.

L. W. Jensen. Upgrade of the railway line between Odense and Aarhus (Opgradering af jernbanen mellem Odense og Aarhus). Master's thesis, Technical University of Denmark, 2012.

L. W. Jensen. Upgrade of the railway between Odense and Aarhus (Opgradering af jernbanen mellem Odense og Aarhus). *Selected Proceedings from the Annual Transport Conference at Aalborg University*, 2013.

L. W. Jensen. An optimisation framework for determination of capacity in railway networks. In *Proceedings of Conference on Advanced Systems in Public Transport (CASPT)*, Rotterdam, 2015.

L. W. Jensen and A. Landex. Measuring Robustness of Timetables at Stations using a Probability Distribution. In *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis*, Copenhagen, 2013a.

L. W. Jensen and A. Landex. Network effects in railway timetables (Netværkseffekter i jernbanekøreplaner). In *Proceedings from the Annual Transport Conference at Aalborg University*, Aalborg, Denmark, 2013b.

L. W. Jensen, A. Landex, and O. A. Nielsen. Evaluation of Robustness Indicators using Railway Operation Simulation. In C. Brebbia, N. Tomii, P. Tzieropoulos, and J. Mera, editors, *Computers in Railways XIV*, pages 329–339, Rome, 2014. WITpress. doi: 10.2495/CR140271.

L. W. Jensen, A. Landex, and O. A. Nielsen. Assessment of Stochastic Capacity Consumption in Railway Networks. In *Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis*, Tokyo, 2015.

L. W. Jensen, A. Landex, O. A. Nielsen, L. G. Kroon, and M. Schmidt. Strategic assessment of capacity consumption in railway networks: Framework and model. *Submitted to: Transportation Research: Part C*, 2016.

R. Jensen and M.-B. Rasmussen. Upgrade of Regional Railway Lines. Master's thesis, Technical University of Denmark, 2015.

A. Kaas. Development and practical use of a capacity model for railway networks. In C. A. Brebbia, editor, *Proceedings of the Conference on Structural Integrity and Passenger Safety*, page 73, Great Britain, 1998a. WITpress.

A. H. Kaas. *Methods to calculate capacity of railways (Metoder til beregning af jernbanekapacitet)*. PhD thesis, Technical University of Denmark, 1998b.

S. G. Klabes. *Algorithmic Railway Capacity Allocation in a Competitive European Railway Market*. PhD thesis, RWTH Aachen, 2010.

D. E. Knuth. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley Professional, 3rd editio edition, 1997. ISBN 0785342896848.

D. E. Knuth. *The Art of Computer Programming, Volume 4, Fascicle 3: Generating All Combinations and Partitions*. Addison-Wesley Professional, 2005. ISBN 0785342853940.

H. Koutsopoulos and Z. Wang. Simulation of Urban Rail Operations: Application Framework. *Transportation Research Record: Journal of the Transportation Research Board*, 2006:84–91, jan 2007. ISSN 0361-1981. doi: 10.3141/2006-10.

L. Kroon, D. Huisman, and G. Maróti. *Optimisation Models for Railway Timetabling*, volume 1, chapter 8, pages 135–154. Eurailpress, Hamburg, Germany, 1st edition, 2008a.

L. Kroon, G. Maróti, M. Helmrich, M. Vromans, and R. Dekker. Stochastic improvement of cyclic railway timetables. *Transportation Research Part B: Methodological*, 42(6):553 – 570, 2008b. doi: 10.1016/j.trb.2007.11.002.

L. G. Kroon, R. Dekker, and M. J. C. M. Vromans. Cyclic Railway Timetabling: A Stochastic Optimization Approach. In F. Geraets, L. Kroon, A. Schoebel, D. Wagner, and C. D. Zaroliagis, editors, *Algorithmic Methods for Railway Optimization*, pages 41–66. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-74247-0. doi: 10.1007/978-3-540-74247-0{_}2.

H. Krueger. Parametric modeling in rail capacity planning. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 31st Conference on Winter simulation*, volume 2, pages 1194–1200. ACM Press New York, NY, USA, 1999. ISBN ISBN:0-7803-5780-9. doi: 10.1109/WSC.1999.816840.

Y.-C. Lai, Y.-H. Liu, and Y.-J. Lin. Standardization of capacity unit for headway-based rail capacity analysis. *Transportation Research Part C: Emerging Technologies*, 57: 68–84, Aug. 2015. ISSN 0968090X. doi: 10.1016/j.trc.2015.05.013.

A. Landex. *Methods to estimate railway capacity and passenger delays*. PhD thesis, Lyngby, Denmark, 2008a.

A. Landex. Evaluation of Railway Networks with Single Track Operation Using the UIC 406 Capacity Method. *Networks and Spatial Economics*, 9(1):7–23, nov 2008b. ISSN 1566-113X. doi: 10.1007/s11067-008-9090-7.

A. Landex. Rail Traffic Engineering - In progress: Fall 2013 - version 0.1. Technical report, Department of Transport, Technical University of Denmark, Lyngby, 2013.

A. Landex and L. W. Jensen. Measures for track complexity and robustness of operation at stations. *Journal of Rail Transport Planning & Management*, 3(1):22–35, 2013.

A. Landex, A. H. Kaas, and S. Hansen. Railway Operation. Technical report, Department of Transport, Technical University of Denmark, Lyngby, 2006.

C. Liebchen, M. Schachtebeck, A. Schöbel, S. Stiller, and A. Prigge. Computing delay resistant railway timetables. *Computers & Operations Research*, 37(5):857 – 868, 2010. doi: http://dx.doi.org/10.1016/j.cor.2009.03.022. Disruption Management.

A. Lindfeldt. *Railway capacity analysis*. PhD thesis, KTH Royal Institute of Technology, 2015.

T. Lindner. Applicability of the analytical uic code 406 compression method for evaluating line and station capacity. *Journal of Rail Transport Planning & Management*, 1(1):49–57, 2011.

R. M. Lusby, J. Larsen, M. Ehrgott, and D. Ryan. Railway track allocation: Models and methods. *OR Spectrum*, 33(4):843–883, dec 2011. ISSN 01716468. doi: 10.1007/s00291-009-0189-0.

L.-G. Mattsson and E. Jenelius. Vulnerability and resilience of transport systems Ű A discussion of recent research. *Transportation Research Part A: Policy and Practice*, 81:16–34, nov 2015. ISSN 09658564. doi: 10.1016/j.tra.2015.06.002.

L. Mussone and R. Wolfler Calvo. An analytical approach to calculate the capacity of a railway system. *European Journal of Operational Research*, 228(1):11–23, 2013. doi: 10.1016/j.ejor.2012.12.027.

A. Nash and D. Huerlimann. Railroad simulation using OpenTrack. In J. Allan, C. A. Brebbia, R. J. Hill, G. Sciutto, and S. Sone, editors, *Proceedings of the 9th International Conference on Computers in Railways*, pages 45–54, Great Britain, 2004. WITpress.

J. Pachl. *Railway Operation and Control*. VTD Rail Publishing, USA, 2002. ISBN 0-9719915-1-0.

J. Pachl. *Timetable Design Principles*, pages 9–42. Railway Timetable & Traffic. Eurail Press, Hamburg, Germany, 1st edition, 2008. ISBN 978-3-7771-0371-6.

J. Parbo. *Public transport optimisation emphasising passengers' travel behaviour*. PhD thesis, Technical University of Denmark, 2015.

M. Piester and M. Thorhauge. Samfundsøkonomiske fordele i køreplaner ved hjælp af passagerforsinkelsesmodeller. Master's thesis, Technical University of Denmark, 2010.

G. Potthoff. *Verkehrsströmungslehre I - Die zugfolge auf Strecken und in Bahnhöfen*. Transpress, Berlin, Germany, 1962.

E. Quaglietta. *A microscopic simulation model for supporting the design of railway systems: development and applications*. PhD thesis, Università degli Studi di Napoli Federico II, 2011.

T. Richter. Data aggregation for detailed analysis of train delays. In *Thirteenth International Conference on Design and Operation in Railway Engineering (COMPRAIL 2012)*, 2012. ISBN 9781845646165.

M. A. Salido, F. Barber, and L. Ingolotti. Robustness for a single railway line: Analytical and simulation methods. *Expert Systems with Applications*, 39(18):13305–13327, dec 2012. ISSN 09574174. doi: 10.1016/j.eswa.2012.05.071.

M. K. Sameni. *Railway Track Capacity: Measuring and Managing*. PhD thesis, University of Southampton, 2012.

B. Schittenhelm. Planning with Timetable Supplements in Railway Timetables. In *Proceedings from the Annual Transport Conference at Aalborg University*, Aalborg, 2011. Aalborg University.

B. Schittenhelm and A. Landex. Development and application of danish key performance indicators for railway timetables. In *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis*, 2013.

B. H. Schittenhelm. *Quantitative Methods for Assessment of Railway Timetables*. PhD thesis, 2013.

T. Schlechte, R. Borndörfer, B. Erol, T. Graffagnino, and E. Swarat. Micro-macro transformation of railway networks. *Journal of Rail Transport Planning & Management*, 1 (1):38 – 48, 2011. doi: 10.1016/j.jrtpm.2011.09.001.

A. Schöbel. Line planning in public transportation: models and methods. *OR Spectrum*, 34(3):491–510, may 2011. ISSN 0171-6468. doi: 10.1007/s00291-011-0251-6.

A. Schöbel and A. Kratz. A Bicriteria Approach for Robust Timetabling. In R. K. Ahuja, R. H. Möhring, and C. D. Zaroliagis, editors, *Robust and Online Large-Scale Optimization*, pages 119–144. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-642-05465-5_5.

W. Schwanhäußer. *Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn.* PhD thesis, Verkehrswissenschaftliches Institut der Rheinisch-Westfälischen Technischen Hochschule, 1974.

R. Sedgewick. Permutation Generation Methods. *ACM Computing Surveys*, 9(2):137–164, June 1977. ISSN 03600300. doi: 10.1145/356689.356692.

Sedgewick, Robert. *Algorithms.* Addison-Wesley Professional, 4th edition, 2011. ISBN 978-0321573513.

T. Siefer. *Simulation*, volume 1 of *Railway Timetable & Traffic*, chapter 9, pages 155–169. Eurailpress, Hamburg, Germany, 1st edition, 2008.

G. Theeg and S. Vlasenko. *Railway Signalling & Interlocking: International Compendium.* 2009. ISBN 978-3777103945.

UIC. Links between railway infrastructure capacity and the quality of operations (UIC Code 405 OR). Technical report, International Union of Railways (UIC), 1996.

UIC. UIC Code 451-1: Timetable recovery margins to guarantee timekeeping. Technical report, International Union of Railways (UIC), 2000.

UIC. Capacity Management (Capman Phase 3). Technical report, International Union of Railways (UIC), 2004a.

UIC. Capacity (UIC Code 406) 1st edition. Technical report, International Union of Railways (UIC), 2004b.

UIC. Capacity (UIC Code 406) 2nd edition. Technical report, International Union of Railways (UIC), 2013.

N. van Oort. *Service reliability and urban public transport design.* PhD thesis, Delft, the Netherlands, 2011.

M. J. C. M. Vromans. *Reliability of Railway Systems.* PhD thesis, Erasmus University Rotterdam, Rotterdam, the Netherlands, 2005.

E. Wendler. The scheduled waiting time on railway lines. *Transportation Research Part B: Methodological*, 41(2):148 – 158, 2007. doi: 10.1016/j.trb.2006.02.009.

J. Yuan. *Stochastic modelling of train delays and delay propagation in stations.* PhD thesis, 2006.

J. Yuan. Capturing stochastic variations of train event times and process times with goodness-of-fit tests. pages 1–12, 2007.

J. Yuan and I. A. Hansen. Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B: Methodological*, 41:202–217, 2007. ISSN 01912615. doi: 10.1016/j.trb.2006.02.004.

Appendices

A Estimation of train running times

As described in sections 2.2 and 2.3, it is necessary to estimate the running times of trains to produce a timetable or model blocking times. For a precise calculation of train running times the following input has to be included:

- Infrastructure (structured in a microscopic model)
  - Allowable speed based on rolling stock type (axle load, tilting train)
  - Gradients
  - Curve resistance (usually neglected)
- Dynamics of rolling stock
  - Mass
  - Length
  - Traction force
  - Air resistance
  - Rolling resistance

A train run may be divided into sections of acceleration, deceleration, coasting and cruising depending on the speed restrictions along the route of the train. In the acceleration phase the resulting force given by the maximum traction of train minus the resistance forces provide the acceleration of the train. This can be solved by using Newton's second law. As the traction force of a train and part of the resistance forces depends on the speed, Newton's second law for estimating the acceleration has to be formulated as a differential equation:

$$F_T(v) - F_R(v) = f \cdot m \cdot \frac{dv}{dt} \tag{A.1}$$

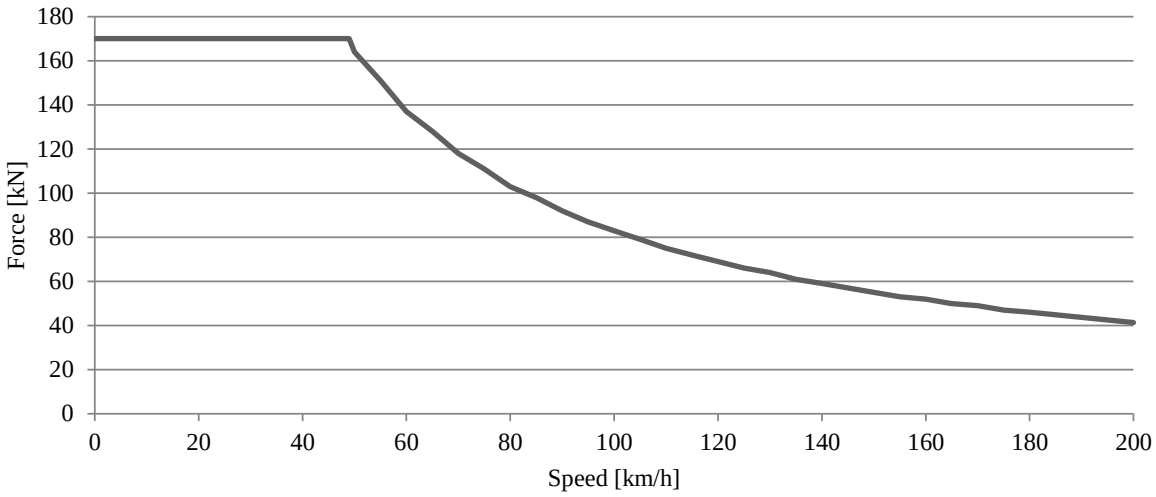


Figure A.1: Traction force as a function of speed for electrical train set (Danish litra ET).

Where $F_{T(v)}$ is the traction force at the wheel rim at speed v, $F_{R(v)}$ is the sum of the resistances as a function of v, m is the mass of the train, f is a constant related to the rotational mass of the train (Brünger and Dahlhaus, 2008) and $\frac{dv}{dt}$ is the acceleration. In figure A.1 the traction effort is shown for an electrical trainset (EMU) as a function of speed.

The deceleration phase may calculated in a similar fashion as the acceleration. However, as most signalling systems calculate the required braking distance of a train based on a constant deceleration, or a series of constant decelerations, this is also the common way to calculate the braking phase in running time estimations (Brünger and Dahlhaus, 2008). The formula for constant deceleration (acceleration), r_b, is given by:

$$s = \frac{1}{2} \cdot r_b \cdot t^2 \tag{A.2}$$

Where the braking acceleration may either be given as a fixed value given by operational rules or estimated for the particular type of rolling stock. The so-called Minden formula can be used to calculate the maximum braking based on the braking percentage, C, of the train (with a 20% supplement):

$$r_b = \frac{6.1 \cdot C + 61}{1200} \tag{A.3}$$

To find the service braking acceleration, the calculated maximum braking acceleration can be adjusted by a factor. For the Danish signalling system a factor of 0.70 is used

(Landex et al., 2006), while Brünger and Dahlhaus (2008) states that a value of 0.75 is often used.

In the cruising phase, the train is travelling at constant speed if the traction effort of the train is able to balance the resistance forces i.e. a resulting force of 0 kN. At very high speeds the resistance forces are very high, especially the (air) drag on the train as the drag increases squared with the speed. Thus cruising at high speeds requires a significant traction effort. The running time in the cruising phase is simply calculated using:

$$t = \frac{s}{v} \tag{A.4}$$

The coasting phase can be calculated similarly to the acceleration phase, with the exception that the traction of the train is not active resulting in a deceleration caused by the resistance forces. Thus the following differential equation has to be solved:

$$F_R(v) = -f \cdot m \cdot \frac{dv}{dt} \tag{A.5}$$

For methods to solve the differential equations, modelling of resistances and an procedure to determine start of acceleration and deceleration phases, see Brünger and Dahlhaus (2008).

For planning purposes, one seeks to find the minimum running time possible. This implies that only acceleration, deceleration and cruising phases are modelled as coasting phases will increase the running time. In practice the train driver will rarely be able to obtain the minimum running time and a supplement is therefore added to account for variance in driver behaviour as well as a measure to absorb or reduce delay (see section 3.1 for more on delay and supplements). The supplement may added by reducing the maximum speed of the train.

However, a train operator will always seek to reduce cost where possible. This has spurred the creation of driver advisor systems (or driver support systems) to aid the train drivers in driving as energy-efficient as possible. E.g. a train is on time and is scheduled to arrive at time t_a at the next station, the minimum running time is t_{min} and the supplement is t_{supp}. The goal is now to arrive on time t_a using as little energy as possible. This can be done by adding a coasting phase before the final braking phase where the running time supplement, t_{supp}, is used to preserve energy. Furthermore, driver advisory systems can be used to reduce delay propagation from one train to the next by ensuring that the second train does not run too close to the first train, thus risking a stop due to a red signal (Albrecht, 2008).

As described earlier, a supplement is added to the minimum running time to account for variance. Between timing points it is up to the train driver to distribute the supplement. Thus there can be a larger variation in how the train is driven between two timing points from driver to driver. This is not desirable when calculating minimum headway times as these depend on the running time in the different block sections on the line (see section 2.3). However, with the rise of driver advisory systems there is a potential to use these systems in running time calculations for timetabling as they encourage consistent and thus more predictable behaviour. Ideally the train is driven automatically yielding absolute predictable driving behaviour. For more information on energy-efficient driving see Albrecht (2008).

Approximation model

In Jensen (2012, 2013) an approximation model for running time calculations was developed called *RTCalc*. This model has been used in this PhD project and developed further to cope with calculations of very complex speed profiles fast and efficiently. The purpose of the model is to calculate the minimum running time to be used for infrastructure planning and draft timetables.

Instead of solving the differential equation for acceleration (eq. A.2) an approximation formula can be used as given by Landex (2013):

$$a_t = a_0 \cdot e^{\frac{a_0 \cdot t}{v_{max}}} \tag{A.6}$$

Where a_o is the start acceleration of the train and v_{max} the maximum (theoretical) speed of the train, G is the gradient and g is the gravitational pull. It should be noted that to calculate the speed as a function of the distance, numerical methods (e.g. Newton-Raphson) has to be used as there is no analytical expression for this. For the braking and cruising stages eq. A.3 and A.4 are used. Curve resistances are not considered.

In algorithm 4 the overall procedure for determining start and end points for acceleration and deceleration is given.

It should be noted that the algorithm can easily be adjusted to use acceleration estimations by solving the differential eq. A.2 instead of the approximation formula in eq. A.6. This can be done by simply replacing the function that calculates acceleration distances and times. This will yield slightly more precise results at the cost of increased calculation time.

Algorithm 4: Pseudocode for placing acceleration and deceleration phases.

```
Data: Gradients, speed profile (allowed speed), train dynamics
Result: Train trajectory yielding the minimum running time
Load gradients and speed profile
Add acceleration/deceleration start and end points for all changes in the speed
 profile
Add acceleration/deceleration start and end points for all stations
while overlapping acceleration points do
 | Shorten, delete or merge acceleration points
end
while overlapping deceleration points do
 | Shorten, delete or merge deceleration points
end
while intersecting/overlapping accelerations and decelerations points do
 | if intersecting then
 |  | Determine intersect speed and shorten acceleration and deceleration
 | else
 |  | Acceleration or deceleration overlaps, delete one
 | end
end
Add running time supplement
return Trajectory (speed as function of distance and time as function of distance)
```

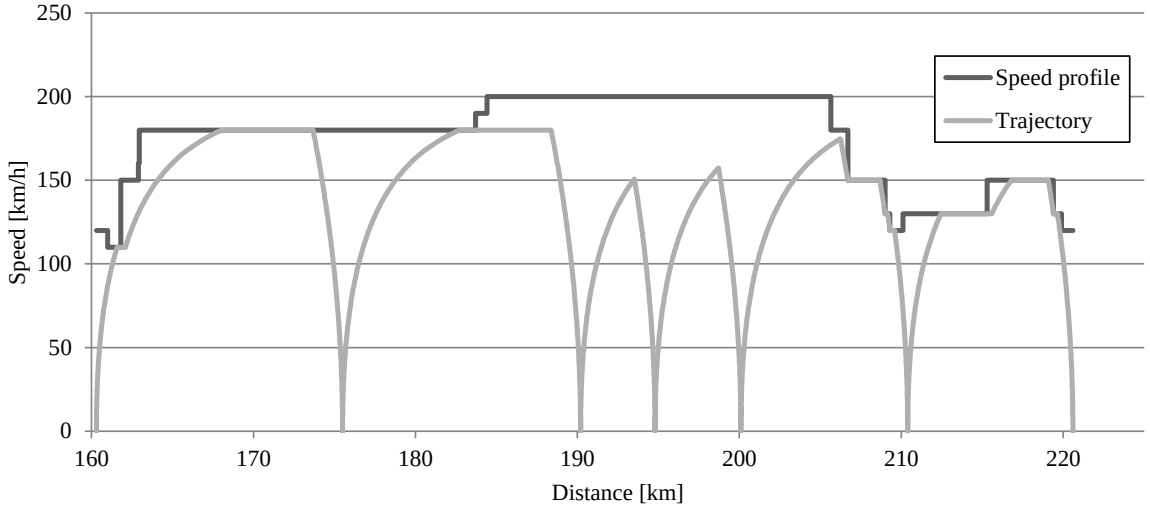


Figure A.2: Speed profile and trajectory estimated for minimum running time.

Jensen (2012) compared the first version of the model with a running time estimation done by using the commercial software RailSys, which uses a differential equation to solve acceleration phase. For three different types of rolling stock the deviation was found to be -1.4%, -0.8% and -1.7% compared to RailSys calculations.

Figure A.2 shows a running time estimation done using the model.

B Flowcharts and pseudo code for the capacity consumption model

This appendix contains a more detailed flowchart on the model of chapter 7 as well as pseudocode on selected algorithms referred to in the chapter. The appendix thus serves as complementary to the mentioned chapter and should therefore be read in conjunction with this.

B.1 Flowchart for capacity consumption model

In section 7.3, procedures and algorithms are described that together form the capacity consumption model. In figure B.1, a flowchart illustrates this model and each procedure of it as well as the input and output for each procedure. For an in depth description of each procedure, see the relevant subsection in section 7.3.

B.2 FIFO network scheduler

The pseudocode in algorithm 5 describes the network scheduler developed by Jensen et al. (2015).

The algorithm shown in algorithm 5 is an asynchronous scheduler for a set of train runs in network using the first-in first-out principle (FIFO). Essentially, for each run the bounding link (bottleneck) is found along this train run's route in the network. The headway to the previous train on each edge in the sequence is then adjusted to the earliest arrival time on the bounding edge. In algorithm 5, the functions, $get\_earliest\_edge\_begin(s, e)$ and $get\_earliest\_edge\_end(s, e)$ determine the earliest begin and end times for a run, $s \in S$, on an edge, $e \in E$, according to edge headways. $run\_edge\_begin(s, e)$ and $run\_edge\_end(s, e)$ are the values of the current begin and end times, respectively, for each run, $s \in S$, on each edge, $e \in E$. These two values are initially set to the earliest possible based on edge

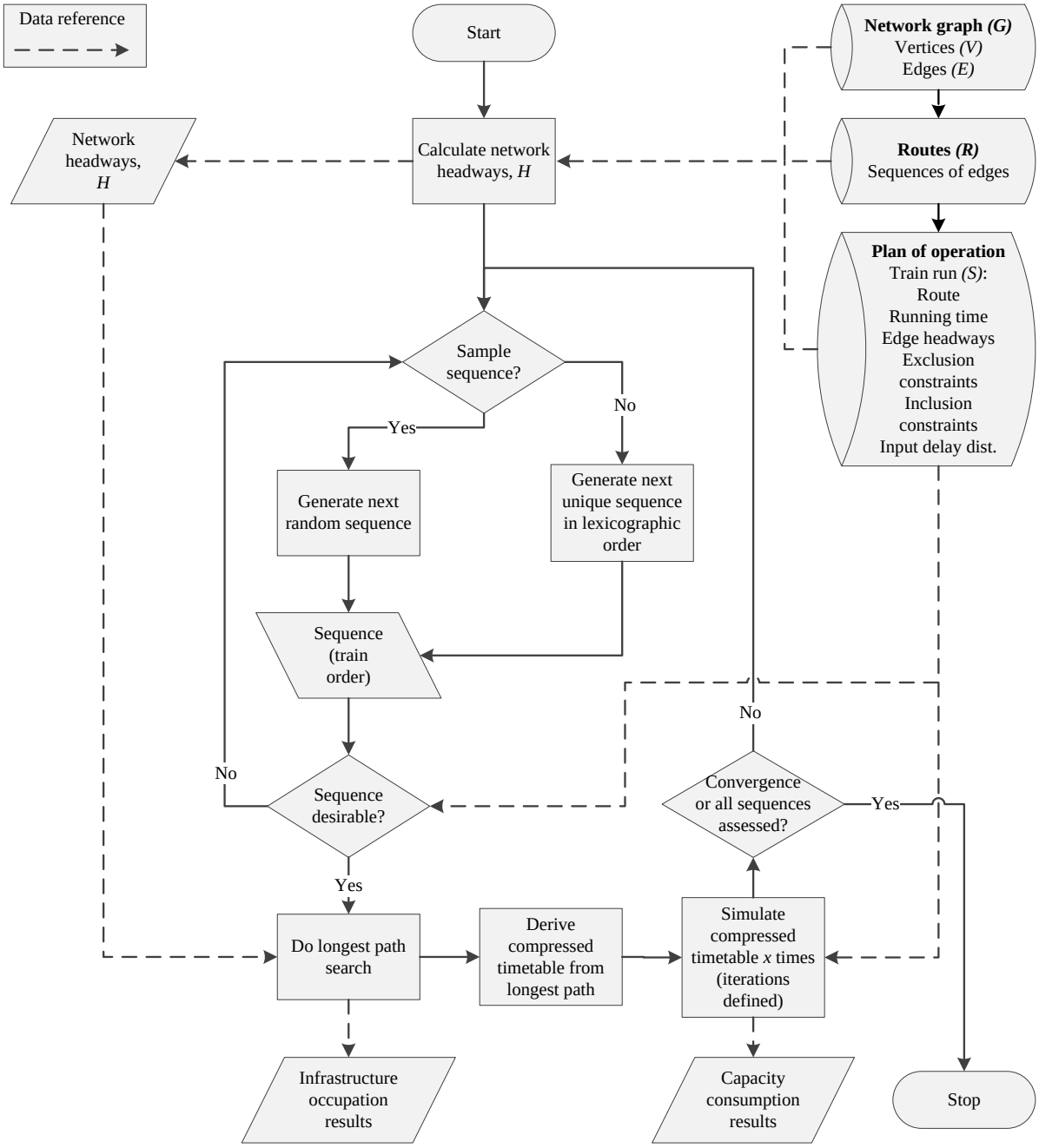


Figure B.1: Flowchart illustrating the model described in section 7.3.

headway constraints only. The algorithm has a computational complexity of $O(|E| \cdot |S|)$ due to the two levels of nested loops in the scheduler.

The scheduler may be used to schedule any number of trains in a network. However, it may also be used to calculate network headways efficiently as done in chapter 7. This can be done by considering two trains at a time and subsequently deduce the network headway between these two trains as the difference in earliest departure times calculated by algorithm 5.

Algorithm 5: First-in first-out (FIFO) scheduler for scheduling train runs in networks.

Data: A network of edges $p \in P$
Z a (sub)sequence of train runs (S)
Routes, R, with edges, E_r in p
Result: $p' = (s, t_0, t_1)$ earliest begin and end times (t_0 and t_1) for every run ($s \in S$) in the network, p
for $i \leftarrow 1$ **to** $|Z|$ **do**
  // Edge sequence of run $Z[i]$ (route)
  $c \leftarrow run\_edge\_sequence(Z[i], p)$
  $bound\_edge \leftarrow \emptyset$, $bound\_edge\_found \leftarrow false$
  /* Initially set begin and end times for current, i, on all edges in c according to edge headway constraints. */
  $run\_edge\_begin(Z[i], c) \leftarrow get\_earliest\_edge\_begin(Z[i], c)$
  $run\_edge\_end(Z[i], c) \leftarrow get\_earliest\_edge\_end(Z[i], c)$
  /* Find bounding edge (bottleneck) in edge sequence, c. First edge, $j = 1$ is initially the bottleneck. */
  for $j \leftarrow 2$ **to** $|c|$ **do**
    if $run\_edge\_end(Z[i], c[j-1]) < run\_edge\_begin(Z[i], c[j]))$ **then**
      $bound\_edge \leftarrow j$, $bound\_edge\_found \leftarrow true$
    else
      $run\_edge\_begin(Z[i], c[j]) \leftarrow run\_edge\_end(Z[i], c[j-1]))$
    end
  end
  /* Offset run begin and end times in edge sequence, c, up till the bounding edge */
  if $bound\_edge\_found$ **then**
    for $k \leftarrow bound\_edge$ **to** 1 **do**
      $run\_edge\_end(Z[i], c[k-1]) \leftarrow run\_edge\_begin(Z[i], c[k]))$
    end
  end
  $p' \leftarrow (Z[i], run\_edge\_begin(Z[i], c[1]), run\_edge\_end(Z[i], c[|c|])$
end

B.3 Longest path search

Algorithm 6 shows pseudo-code for a longest path search in a directed acyclic graph (DAG) using dynamic programming. The algorithm can be used to calculate infrastructure occupation for an acyclic train sequence represented by a DAG (denoted minimum headway graph). In the graph edges represent train dependencies to subsequent trains in the sequence with weights equal to the minimum required headway times. The infrastructure occupation is the length of the longest path found.

Algorithm 6: LPS: Longest path search in DAG

Data: Minimum headway graph $G_h = (V_h, E_h)$ with edge weights (minimum headways), h_{ij}, and topological sorting (train sequence), Z, of G_h
Result: Longest path to all vertices, Z', longest connected path, Z'_c, and start times, $t_z \in T_z$

```
for each vertex z in the topological order Z do
    t_z ← 0
    for each vertex v connected to z with an incoming edge e_vz do
        if t_v + h_vz > t_z then
            t_z ← t_v + h_vz
            e_max ← e_vz // Record edge
        end
    end
    Add e_max to Z'
end
Z'_c ← Z'
Remove all edges from Z'_c where the target vertex has outdegree 0 in Z'_c
return Z', Z'_c and T_z
```

DTU Transport performs research and provides education on traffic and transport planning. It advises the Danish Ministry of Transport on infrastructure, economic appraisals, transport policy and road safety and collects data on the transport habits of the population. DTU Transport collaborates with companies on such topics as logistics, public transport and intelligent transport systems.

DTU Transport
Department of Transport
Technical University of Denmark

Bygningstorvet 116B
DK-2800 Kgs. Lyngby
Tel. +45 45 25 65 00
Fax +45 45 93 65 33

www.transport.dtu.dk