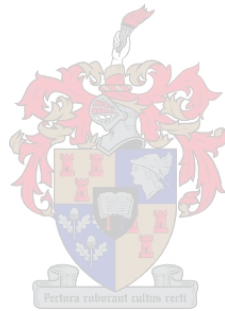


Modelling service reliability of a heterogeneous train fleet operating on aged infrastructure

Nathan Wilson

**Thesis presented in fulfilment of the requirements for the degree of
Master of Engineering (Industrial Engineering) in the Faculty of Engineering at Stellenbosch
University**



Prof Romano Del Mistro, Prof Cornelius J. Fourie, Prof Corne Schutte

March 2017

The financial assistance of the PRASA Engineering Research Chair towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to PRASA

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2017

Copyright © 2017 Stellenbosch University

All rights reserved

Abstract

The Passenger Rail Agency of South Africa is in the process of introducing new rolling stock into their current aged fleet of rolling stock. This poses several technical challenges relating to amongst other the operation of the mix of old and new trains on the same infrastructure. The objective of this study was therefore to determine the effect on service reliability in terms of punctuality, when old trains are incrementally replaced by new trains. Punctuality was measured by number of delays, total delay minutes and average delay duration over a specified time period.

A discrete-event simulation model was developed using *Anylogic* simulation software. The line between Chris Hani and Cape Town stations on the Western Cape Metrorail network was chosen as case study for the model. Two cases were modelled with each consisting of 14 scenarios. Case 1 assumed no reliability improvement to the overall rail system. Since the specific route consisted of 14 trains shuttling to and from Cape Town, each scenario represented the replacement of an old train with a new train until the whole fleet consisted of only new trains. Case 2 modelled the same scenarios, except it was assumed that the system's reliability was improved by an arbitrary value of 50%.

In Case 1 a 29% improvement in number of delays, 37% improvement in total delay minutes, and 11% improvement in average delay duration were seen when Scenario 0 (base case) was compared to Scenario 14 (future case with all 14 old trains replaced). In Case 2 a 31% improvement in number of delays, 36% improvement in total delay minutes, and 7% improvement in average delay duration were seen.

When Case 1 and 2 are compared on a scenario for scenario basis (e.g. Case 1, Scenario 0 compared to Case 2, Scenario 0) it was found that the 50% reliability improvement of the overall system resulted in an average improvement of 13% in number of delays, 19% in total delay minutes, and 6% in average delay duration. The overall improvement from zero new trains and no system reliability improvement (Case 1, Scenario 0) to 14 new trains and 50% system reliability improvement (Case 2, Scenario 14) resulted in a 39% reduction in number of delays, 47% reduction in total minutes delay, and 13% reduction in average delay duration.

The model therefore shows how a train service can improve in terms of punctuality, when reliability improvements are made such as new rolling stock or overall system improvements that resolve primary delay causes. The findings of this study can therefore be used to support decisions related to capital investments into reliability improvements and new rolling stock commissioning strategies.

Opsomming

Die Passasier Spoor-agentskap van Suid-Afrika is huidige in die proses om nuwe rollende materiaal in te faseer in die huidige vloot van rollende materiaal. Hierdie proses skep verskeie tegniese uitdagings ten opsigte van die bedryf van die mengsel van ou en nuwe treine op dieselfde infrastruktuur. Die doel van hierdie studie was om te bepaal wat die effek op diensbetroubaarheid in terme van stiptelikheid is, wanneer ou treine inkrementeel vervang word met nuwe treine. Stiptelikheid was gemeet deur hoeveelheid vertragings, totale vertragingsminute, en gemiddelde vertragingsduur oor 'n gespesifiseerde tydperk.

'n Diskrete-gebeurtenis simulasiemodel was ontwikkel met die gebruik van *Anylogic* sagteware. Die spoorlyn tussen Chris Hani- en Kaapstadstasie op die Wes-Kaapse Metrorail netwerk was gekies as gevallestudie vir die model. Twee gevalle was gemodelleer met elkeen wat bestaan uit 14 scenarios. Geval 1 het aangeneem dat geen betroubaarheidsverbeteringe aan die oorhoofse spoorwegsisteem aangebring was nie. Aangesien dié spesifieke roete 14 treine bevat wat na en van Kaapstad reis, stel elke scenario die inkrementele vervanging van 'n ou trein met 'n nuwe trein voor totdat die hele vloot uit slegs nuwe treine bestaan. Geval 2 het dieselfde scenarios gemodelleer, behalwe dat 'n aanname gemaak was dat die betroubaarheid van die oorhoofse sisteem met 50% verbeter was.

In Geval 1 was 29% verbetering in hoeveelheid vertragings, 37% verbetering in totale vertragingsminute, en 11% verbetering in gemiddelde vertragingsduur gevind. In Geval 2 was 31% verbetering in hoeveelheid vertragings, 36% verbetering in totale vertragingsminute, en 7% verbetering in gemiddelde vertragingsduur gevind.

Wanneer Geval 1 en 2 met mekaar vergelyk word op 'n scenario-teen-teen-basis, was daar gevind dat die 50% betroubaarheidsverbetering aan die oorhoofse sisteem gelei het tot 'n gemiddelde verbetering van 13% in hoeveelheid vertragings, 19% verbetering in totale vertragingsminute, en 6% verbetering in gemiddelde vertragingsduur. Die algemene verbetering vanaf geen nuwe treine en geen sisteem-betroubaarheidsverbetering tot 14 nuwe treine en 50% sisteemverbetering het gelei tot 39% verbetering in hoeveelheid vertragings, 47% verbetering in totale vertragingsminute, en 13% verbetering in gemiddelde vertragingsduur.

Die model wys dus hoe 'n treindiens kan verbeter in terme van stiptelikheid wanneer diensbetroubaarheid verbeteringe aangebring word soos nuwe rollende materiaal en oorhoofse sisteme verbetering wat primêre vertragings verminder. Die bevindings van die studie kan daarom gebruik word om besluitneemings te ondersteun met verband tot kapitale investerings in diensbetroubaarheids verbeteringe en rollende materiaal inbedryfstelling strategieë.

Acknowledgements

I would hereby like to acknowledge the honest, fervent help and guidance from my study leader, Prof Del Mistro. Thank you for the time and effort you committed to guide and review this work.

Likewise I would like to say thank to Prof Fourie for his support and guidance throughout the duration of this study. Thank you also for providing the platform, structure and funding of the PRASA Research Chair within which I could base my study. Furthermore I would also like to extend my gratitude to Pieter Conradie and Olabanji Asekun for their support.

I would also like to thank my parents Gous and Lisa Wilson for their continuous support and prayer and for inspiring me to commence on this journey.

Lastly I would like to give a special thanks to my grandparents, Hannes and Isabel Venter for helping on the language review of this document.

Dedications

I would like to dedicate this work to my Father in Heaven.

“Therefore, whether you eat or drink or whatever you do, do all to the glory of God.”

- 1 Corinthians 10:31

Table of Contents

Declaration	ii
Abstract	iii
Opsomming	iv
Acknowledgements	v
Dedications	vi
List of Figures	x
List of Tables.....	xii
List of Abbreviations.....	xiii
1. Introduction	2
1.1. Background	2
1.2. Problem Statement	2
1.3. Brief Chapter overview	3
2. Case study background.....	4
2.1 Overview	4
2.1.1 Asset base	4
2.1.2 Operations	5
2.2 Stakeholders' motivation for renewal	5
2.2.1 PRASA strategy	5
2.2.2 Railway Safety Regulator.....	6
2.2.3 Ministry of Transport	7
2.2.4 Summary	8
2.3 Modernization program	8
2.4 Train types.....	9
2.5 Summary	9
3. Literature Review	11
3.1 Introduction	11
3.2 Mathematical models and heuristics algorithms	11
3.2.1 Queueing models.....	11

3.2.2	Job shop models	15
3.2.3	Tabu search.....	17
3.2.4	Genetic Algorithm.....	17
3.3	Simulation models.....	18
3.3.1	Macroscopic simulation models and software.....	18
3.3.2	Microscopic simulation models and software	22
3.4	Train system punctuality and reliability	24
3.5	Summary	25
4.	Model development.....	26
4.1	Basic outline.....	26
4.1.1	Outputs	26
4.1.2	Inputs	27
4.1.3	Simulation software.....	32
4.1.4	Summary	33
4.2	Model – Infrastructure sub-model.....	34
4.3	Model – System agent	35
4.4	Model – Train agent	38
4.5	Limitations and assumptions	39
4.5.1	Simulating delays	39
4.5.2	Acceleration and deceleration properties	40
4.5.3	Train passing	41
4.5.4	Peak and off-peak delay events	42
4.6	Summary	42
5.	Case study model.....	43
5.1	Inputs.....	44
5.1.1	Perway.....	44
5.1.2	Rolling stock.....	44
5.1.3	Stations	45
5.1.4	Signals	47
5.2	Stochastic inputs.....	47

5.2.1	Location.....	48
5.2.2	Time.....	48
5.2.3	Duration.....	49
5.3	Validation without delays.....	52
5.4	Validation with delays.....	53
5.4.1	Number of delays	57
5.4.2	Total minutes delay	59
5.4.3	Average delay duration.....	60
5.4.4	Conclusion.....	62
6.	Scenarios and model outputs	63
6.1	Overview	63
6.2	Case 1	64
6.3	Case 2	67
6.4	Comparison of Case 1 and Case 2.....	70
6.4.1	Number of delays	70
6.4.2	Total minutes delay	70
6.4.3	Mean delay duration.....	71
7.	Conclusion.....	73
7.1	Case 1	73
7.2	Case 2	74
7.3	Case 1 and Case 2 comparison	74
8.	Recommendations	76
8.1	Further research.....	76
8.2	Introduction of new trains	76
9.	References	77
	Appendix A1 – Timetables and trip times.....	79
	Appendix A2 – Delay causes	81
	Appendix B – Model algorithms	83
	Appendix C – Published article.....	85

List of Figures

Figure 3-1: Simple job shop model	15
Figure 3-2: Small network with 9 block sections and two trains	16
Figure 3-3: Job shop graph of two trains [15]	16
Figure 3-4: Normal timetable without delays [25]	20
Figure 3-5: Simulated timetable diagram with delays [25]	20
Figure 3-6: Simone simulation animation output [26]	21
Figure 3-7: Total knock-on delays at the destination station [25]	21
Figure 4-1: Marvey diagram for normal and delayed homogeneous rail traffic consisting of 5 trains with the primary delay occurring at Signal 2.	30
Figure 4-2: Marvey diagram for normal and delayed homogeneous rail traffic consisting of 5 trains with the primary delay occurring at Signal 7.	31
Figure 4-3: Marvey diagram for normal and delayed rail traffic consisting of 4 slow trains and 1 fast train with the primary delay occurring at Signal 7	32
Figure 4-4: Model outline.....	34
Figure 4-5: Basic Anylogic discrete event model. The top row shows the standard DE process blocks, while the bottom row shows how these were translated to rail infrastructure terms.....	34
Figure 4-6: Process block arrangement to account for two train types	36
Figure 4-7: System agent algorithm accounting only for one train type	37
Figure 4-8: Flow diagram of the Train agent	39
Figure 4-9: Model acceleration curve vs real acceleration curve	41
Figure 5-1: Chris Hani to Cape Town network diagram	43
Figure 5-2: Train agent's speed and acceleration algorithm	45
Figure 5-3: GIS map of the station and signal locations between Chris Hani and Cape Town stations	47
Figure 5-4: Extract from the delay data received from PRASA in Excel format.....	48
Figure 5-5: Distribution of the number of delays and number of trains during each hour of the day, sampled over 6 months only for trains running up (i.e. Chris Hani to Cape Town)	49
Figure 5-6: The observed and estimated cumulative distributions for rolling stock related primary delays.....	52
Figure 5-7: The difference between the scheduled and modelled trip times when the model is run without delays	53
Figure 5-8: Scatter plot of the relationship between primary delays and the resulting sum of delays ..	56
Figure 5-9: Cumulative distributions describing the Number of delays per week from the observed and modelled data sets	58
Figure 5-10: Cumulative distributions describing the Total minutes delay per week from the observed and modelled data sets.....	60

Figure 5-11: Cumulative distributions describing the Average delay duration per week from the observed and modelled data sets	62
Figure 6-1: Number of delays for each department.....	64
Figure 6-2: Total delay minutes for each department	64
Figure 6-3: Number of delays for Scenarios 0-14 and Case 1.....	65
Figure 6-4: Total sum of delays for Scenarios 0-14 and Case 1.....	66
Figure 6-5: Mean delay duration for Scenarios 0-14 and Case 1	67
Figure 6-6: Number of delays for Scenarios 0-14 and Case 2.....	68
Figure 6-7: Total minutes delays for Scenarios 0-14 and Case 2.....	68
Figure 6-8: Relationship between primary delay duration and sum of delays from modelled data	69
Figure 6-9: Mean delay duration for Scenarios 0-14 and Case 2	69
Figure 6-10: Number of delays comparison of Case 1 and Case 2	70
Figure 6-11: Total minutes delay comparison of Case 1 and Case	71
Figure 6-12: Mean delay duration comparison of Case 1 and Case 2.....	71

List of Tables

Table 2-1: Metrorail Western Cape Asset Base	4
Table 2-2: Rail traffic volumes [3].....	6
Table 2-3: The cost of operational occurrences and security related incidents[3]	7
Table 2-4: Passenger numbers [1]	8
Table 4-1: Summary of the different simulation software packages considered for this study	33
Table 5-1: Station inputs	46
Table 5-2: Summary of the primary delays under each department for the 6 month period.....	50
Table 5-3: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for delay durations of rolling stock related delays.....	51
Table 5-4: Last calibration round results.....	54
Table 5-5: Summary of primary delays and resulting sum of delays modelled compared to data.....	55
Table 5-6: Mean and variance values for each parameter calculated from the observed data	57
Table 5-7: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for Number of delays per week.....	57
Table 5-8: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for Total minutes delayed per week	59
Table 5-9: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for Average delay duration per week.....	61
Table 6-1: Summary of results	72

List of Abbreviations

PRASA	Passenger Rail Agency of South Africa
Perway	Permanent way
NEMO	Network Evaluation Model
FCFS	First-Come-First-Serve
LCFS	Last-Come-First-Serve
AHP	Analytical Hierarchy process
DEA	Data Envelopment Analysis
LP	Linear Programming
RTC	Rail Traffic Controller
GPS	Global Positioning System
DE	Discrete Event
PTI	Platform-Train Interchange
RSR	Rail Safety Regulator
CCTV	Closed-Circuit Television
GIS	Geographic Information System
CBD	Central Business District
K-S	Kolmogorov-Smirnoff

1. Introduction

1.1. Background

Railway network companies often have the need to model and simulate the operation of their trains. This need usually arises with the expansion of infrastructure or the addition of new rolling stock and services. Infrastructure expansion entails adding new links, stations, or additional lines. Furthermore permanent way (perway), electrical and signal maintenance all contribute to train operations being disrupted to some extent. Also adding train services or new rolling stock requires major operations planning and rescheduling. Forecasting the effect on the operation of the network before the implementation of such changes is a crucial component to planning. Bottlenecks, line capacities, demand satisfaction and delay propagations are all areas that need to be identified and calculated before large capital amounts are spent. This can be done by the use of mathematical models and simulation.

The Passenger Rail Agency of South Africa set into place a modernization program to renew various infrastructure and rolling stock components of the current rail network in South Africa. Part of the modernization program is also to introduce new rolling stock to the current fleet. This poses various technical as well as socio-economic complexities. This study will focus on the technical complexities, even though the socio-economic factors may carry more weight in terms of the final decision as to how and where the new trains will be deployed. To partly account for the socio-economic agenda of the South African Government, it is anticipated that new trains will be deployed in the most densely populated areas of the network. This study is based on the Western Cape network, and therefore the Chris Hani to Cape Town route was chosen as case study.

There are various technical issues that will have to be addressed before the new trains can be introduced. These issues all relate to the operational readiness of the system in terms of electrical, perway, signalling and service depots. One example of these issues (even though it will not be covered in this study) is temporary speed restrictions caused by poor track condition. To utilise the faster speed characteristics of the new trains, the track has to be fixed and speed restrictions be lifted.

1.2. Problem Statement

In the Western Cape network the current fleet of trains does not meet the peak demand, and therefore in the short to medium term, the new trains will be operated with the old trains instead of simply replacing them. New trains will be introduced into the current fleet as they are rolled out from manufacturing, meaning that the fleet composition of old and new trains will be changed incrementally. Because the new trains have faster speed characteristics and are expected to be more

reliable, it creates a heterogeneous train fleet. The question exists as to how each new train being introduced will affect the overall service in terms of punctuality?

This study thus aims to answer this question by means of a dynamic simulation model of the Chris Hani to Cape Town rail line.

1.3. Brief Chapter overview

Chapter 2 will explain the motivation behind the new rolling stock and modernization of the current rail system. The background of the case study is therefore articulated in terms of the agenda of PRASA and the Ministry of Transport in Chapter 2. Chapter 3 will cover the literature study of mathematical and simulation models of rail networks. It explains how queuing and optimization models were used to solve train disruption and scheduling problems. Chapter 4 will then explain how the model of this study is constructed and elaborates on the assumptions and limitations associated with the approach.

How the model developed in Chapter 4 was used to the model the case study in Chapter 2, is then covered in Chapter 5. Chapter 5 shows how the specific software was used to build the model to produce the desired outputs.

The results of the model are then illustrated and discussed in Chapter 6. Chapter 7 will draw the final conclusions together and discuss the relevance of the findings. Finally, Chapter 8 will list the various recommendations that came from the study and suggest areas of further study in the future.

2. Case study background

In this chapter the background to the case study of the *PRASA* rail line in the Western Cape, South Africa will be described. Section 2.1 will give an overview of who *PRASA* is as an organization and Section 2.2 will explain the motivation and need for modernising the network. Section 2.3 will give an overview of what the modernization program entails and Section 2.4 will cover the different train types which are currently and will in the future operate on the Western Cape network.

2.1 Overview

The entire South African railway is operated by two main state owned companies i.e. *PRASA* and Transnet. *PRASA* is dedicated to only passenger transport, whereas Transnet is responsible for freight transport services. This study however will only focus on *PRASA*.

According to the terms of the Legal Succession SATS Act, the primary goals for *PRASA* are to provide urban rail commuter and long haul passenger services as well as long haul bus services. While providing these services, the secondary objective of *PRASA* is to utilize its acquired assets to generate income. Furthermore, *PRASA*'s responsibilities are "to effectively develop and manage rail and rail related transport infrastructure to provide efficient rail and road based passenger transport within, to and from Urban and Rural areas." [1]

2.1.1 Asset base

The commuter rail network (*Metrorail*) is electrified by 3kV lines where the *Shosholoza Meyl* network consists of 3kV, 25kV and diesel lines. By "diesel lines", it is meant that trains running on those sections are powered by diesel locomotives instead of electric motor coaches.

Metrorail serves four regions i.e. Eastern Cape, Kwazulu-Natal, Gauteng and Western Cape. This study will only look at one line the Western Cape region, and therefore Table 2-1 shows a summary of the Assets of *Metrorail* Western Cape.

Table 2-1: *Metrorail* Western Cape Asset Base

Metrorail Western Cape Asset base	
Stations	123 units
Track	489 km
Reserve	10 400 ha
Turnouts	610 units
Level crossings	70 units
Rail reserve	320 km
Bridges	96 units
Foot bridges	19 units
Culverts	380 units
Sea walls	9 km

2.1.2 Operations

According to PRASA's Annual Report 2012/2013 [1], *PRASA*'s operational units are responsible for the following roles:

- Planning and managing of day-tot-day operations
- Transport service scheduling
- Maintaining infrastructure and rolling stock
- Collecting fare and rental incomes
- Providing passenger security
- Implementing operational safety plans

The Western Cape network is operated in three corridors. The Central Line carries the largest amount of passengers and includes the routes from Cape Town to the Cape Flats, Simons Town and Bellville. The Southern Line is the route between Bellville and Strand and the Northern Line includes the routes connecting the Wellington, Worcester and Malmesbury areas to the Northern suburbs of the Cape. It is estimated that the network covers around 75% of residential areas across six municipalities in the Western Cape [1].

A fleet of 88 trains service an estimated 14.5 million passenger journeys per month with an average punctuality of 78%. Train frequencies vary between 3- and 15 minutes depending on which corridor or route and the passenger volumes [1].

2.2 Stakeholders' motivation for renewal

In this Section the need to modernize the current state of *PRASA* will be motivated. The viewpoints of *PRASA*, the Railway Safety Regulator and the Minister of Transport will be summarized and discussed.

2.2.1 PRASA strategy

PRASA's Annual Report 2012/2013 explains its strategy as follows: "The Strategy of *PRASA* seeks to create a modern public entity by 2017 that would be able to deliver quality passenger services on a more sustainable basis." [1] *PRASA* intends to implement this strategy through capacity investments in modern trains, signalling and telecommunications, infrastructure, new stations, access control and other operating systems. This will then lead to improved service delivery. It also intends to utilize the value of its telecommunication network and property portfolio. *PRASA* has thus set the following goals for its *Metrorail* service:

- Cash generation adequate to cover its operational funding requirements
- The utilization of assets to grow its property portfolio in order to fund future investments
- New stations and facilities
- A public transport share of between 35-40% for rail.

- Modern reliable infrastructure
- Metro train frequencies of 3 – 5 minutes during peak periods.
- Long-distance rail services able to compete with road transport
- Operations meeting necessary quality and safety standards

According to Sifiso Buthelezi [2], Chairman of *PRASA*, the main objective of *PRASA* is to provide quality public transport to connect people from their homes to their work and areas of economic activity. The challenge however is to provide a safe, reliable and predictable service amid the following circumstances:

- Old rolling stock with an average age of 40 years
- Rolling stock shortages
- Outdated signalling system
- Aged infrastructure
- The sabotaging of trains and cables
- Engineering knowledge and skills shortages

These problems are planned to be resolved through the modernisation program discussed Section 2.3.

2.2.2 Railway Safety Regulator

In this Section a summary of the State of Safety Report for the year 2013/2014 will be given. It can be argued that many of the accidents and safety related incidences can be blamed on the outdated and under-invested infrastructure and rolling stock. Table 2-2 shows traffic volumes from the financial years 2008/2009 to 2013/2014. A reduction of 14.7% in passenger numbers for *PRASA* from 2012/2013 to 2013/2014 must be noted. Table 2-3 shows the cost of operational rail occurrences and security-related incidents from the year 2008/2009 to 2013/2014. A significant drop in collisions and derailments can be noted for the year 2013/2014. However, the cost of level crossing accidents increased drastically from R500 000 to R15.3 million, while vandalism and train fires are the largest contributors to the cost of accidents and incidents amounting to a total of R112.1 million.

Table 2-2: Rail traffic volumes [3]

Operator/Year	08/09	09/10	10/11	11/12	12/13	13/14
TFR (Million Train Kilometres)	55.5	49.3	45.9	46.3	46.0	46.9
TFR (Billion Ton Kilometres)	113.0	113.0	117.9	126.5	132.4	134.6
PRASA (Million Passenger Kilometres)	16 874	12 312	12 232	13 651	16 735	14 269
PRASA (Million Train Kilometres)	28.4	30.0	26.3	19.9	24.53	24.97
Gautrain (Million Passenger Kilometres)	None	None	11.5*	119.2	340.8	419.8
Gautrain (Million Train Kilometres)	None	None	0.3*	1.43	4.07	4.4

PRASA reported a total of 8 train fires, one train collision and one station building fire as the top 10 incidents contributing to the costs depicted in Table 2-3 for the financial year of 2013/2014. The train fires are mostly caused by acts of vandalism or protests.

Table 2-3: The cost of operational occurrences and security related incidents[3]

DESCRIPTION	COSTS [R million]					
	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14
Theft & Vandalism	7.6	6.8	15.1	9.6	9.6	17.5
Train Fires	43.5	72.8	64.6	126.2	116.7	94.6
Level Crossing Accidents	0.8	0.4	0.2	0.6	0.5	15.3
Collisions	3.5	21.9	24.4	61	30	2.5
Derailments	5.2	9.7	19.3	6.1	10.1	0.21
GRAND TOTAL	60.7	111.7	123.5	203.2	167.6	130.3

One of the main concerns for *PRASA* in terms of safety is platform-train interchange (PTI) of passengers. The *RSR* reported 83 incidents of passengers falling between the train and the platform and 615 incidents of passengers falling on the platform while entraining and detraining a train. It is reported that overcrowding and reckless behaviour of passengers are the main causes of these incidents. Another important factor to consider is the vertical gap between some of the station platforms and train floors. The study done by the *RSR* showed that stations with a gap of 20cm and more, experienced significantly more incidences than stations with lesser of a gap [3]. However the study concludes that passenger behaviour contributes to between 65 and 75% of PTI occurrences while internationally, PTI occurrences typically amount to 20-25%.

Passenger behaviour that result in PTI occurrences can be related to under-capacity during peak periods. Most incidents occur during the periods 04:00-08:00 and 16:00-20:00. The *RSR* study concludes by saying that trains are not allocated effectively enough to meet passenger demand. Busy lines are thus under-capacity and quieter lines are over-capacity.

2.2.3 Ministry of Transport

The Minister of Transport stated in September 2014 that the key objectives of *PRASA*, since its inception in 2009, are customer centricity, modernization, state-of-the-art technology, efficiency and punctuality [4]. In her speech it was also reported that 500 coaches were out of service due to vandalism and theft. The train punctuality target of 85% was missed by 5% in 2014, and according to the Minister the inability to attain service delivery objectives may be as a result of a lack of capacity. Furthermore the decision to stop *Shosholoz Meyl's* operational subsidy during the 2010/2011 financial year has caused a serious drop in service quality and passenger numbers. Table 2-4 shows the growth and decline of passenger numbers during the years 2012 – 2013. Note that *Shosholoz Meyl* experienced a drop of 11.2%. The Minister stated that financial support is essential for the continuation of the long distance passenger rail service, since the termination thereof may result in severe socio-economic consequences.

Table 2-4: Passenger numbers [1]

	Metrorail	Shosholoza Meyl	Autopax
2011/12	516 392 805	1 423 173	3 163 424
2012/13	528 204 625	1 263 500	3 146 768
Change year on year	2.3%	-11.2%	-0.5%

2.2.4 Summary

It is clear from the *PRASA* strategy that the modernization of rail infrastructure and rolling stock is a priority and a reality. To provide a “quality passenger service on a more sustainable basis”[1] will require major capital investment. The reports of the *RSR* and Minister of Transport reveal the true state of the passenger rail service currently. Inadequate station platforms, over-crowding of trains, insufficient line capacities, passenger train punctuality and reliability are all issues that need drastic action and capital expenditure. To add to the current inefficiencies of the railway, passenger demand for metro services are increasing, and is expected to increase in the future. The planned rail rival can reduce road traffic, fuel reliance and carbon emissions. An effective and reliable metro rail service can also attract investment and increase a region’s economic capacity.

A modernisation program to revive the metro rail infrastructure and rolling stock is therefore a crucial necessity for the development of the South African economy.

2.3 Modernization program

A new rail fleet of 600 new X'Trapolis Mega train sets are planned to be built to replace the existing *Metrorail* fleet. The first 20 sets will however be built in *Alstom*’s Lapa manufacturing plant in São Paulo, Brazil (*Alstom* is the majority shareholder in the *Gibela* Consortium). This will not only ensure that the first batch of trains are built and delivered in a short period, but will also serve as a practical training exercise for *Gibela*’s South African employees. The first coaches are expected to be completed in 2016. *Gibela* will be building a production facility in Dunnottar, South Africa to produce the remaining 580 train sets. It is estimated that the facility will employ around 1500 people and create 8000 indirect jobs [5]. The tender amounts to a total of R123bn over the next 20 years.

According to *PRASA* [6] only 14% of the current signalling systems have not exceeded their design life. The outdated technology contributes to the unreliable service currently experienced by passengers. It is planned to build new train control centres and signalling systems to improve operational safety, capacity and rolling stock performance. The first phase of the project is estimated to cost R7bn and it is anticipated to be completed by June 2020. The renewal will include the regions Gauteng 1, Durban, Western Cape and Gauteng 2.

Five Maintenance depots are also going to be renewed and modernized. These depots are designed to accommodate both the old and new fleet of rolling stock. The aim is to install new cranes and add the function of in-floor lifting. An investment of R1.9bn has been allocated to depot upgrades[6].

A total of 135 stations were prioritized for upgrade. Currently construction is commencing on 14 stations. The new stations will be equipped with speed gates, electronic information displays, public address systems and CCTV. This will improve passenger control and make the metro service more attractive and safe. The estimated expenditure of R1.5bn is expected [6].

PRASA is also planning to upgrade the line speeds from to 120km/h and to 160km/h for the express lines. This will involve track and sleeper replacements, drainage upgrading, ballast screening and realignment of tracks. Overhead traction equipment and substations will also need upgrading to accommodate the faster X'Trapolis trains. An expenditure framework of R1.6bn is expected [6].

New rail links and network expansions are planned to keep up with economic growth in areas such as Bellville. The Blue Downs link for example will move passengers from Phillipi and Khayelitsha directly to Bellville, and relieve the overcrowded route to Cape Town. The other priority links include [6]:

- Atlantis corridor and link
- Phillipi – Southfield link
- Cape Town International Airport link
- Khayelitsha - Somerset West link

These links will improve commuter accessibility of the Western Cape's metro service greatly. There is however still a lot of room for improving the current infrastructure and service.

2.4 Train types

Currently in the Western Cape network, three different types of trains operate namely:

- 5M2A
- 8M
- 10M

The X'Trapolis trains (EM01) will only have the coaches between the head and tail coaches motorized. These motor coaches will each have 4 motorized axles. Module compositions of 4, 5, and 6 can be made to give 50%, 60% and 66% motorized ratios respectively. The EM01 will be able to accelerate at $0,83\text{m/s}^2$ with a top speed of 120km/hr^1 . The seating capacity ranges between 234 and 380 per coach. A 6 coach module which is regarded as the standard module, will thus be able to carry between 1088 and 1218 passengers [7]. It is hoped that these trains will help relieve the high demand experienced in peak periods of the day when train over-crowding is a frequent problem.

2.5 Summary

The modern technology and upgrading projects discussed in this chapter will not be sufficient to replace all whole system and neither will it happen instantly. This means that ways have to be found to incorporate and integrate the new technology with the old technology in such a way to improve service

¹ Compared to $0,35\text{m/s}^2$ and 80km/hr of the 5M2A trains currently most commonly used on the Western Cape network.

delivery effectively and sustainably. This study will therefore focus on how the new EM01's will improve the service in terms of punctuality if they are introduced incrementally.

3. Literature Review

3.1 Introduction

This Chapter will discuss the two spectrums of modelling train networks namely analytical models and simulation models. In Section 3.2 mathematical models and heuristic algorithms will be discussed whereas in Section 3.3, simulation models will be covered. Section 3.4 will then discuss train system punctuality and reliability.

3.2 Mathematical models and heuristics algorithms

Analytical models tend to be limited in scope and complexity, but mostly form the basis on which simulation models are built. With the advances made in computing power capabilities in recent years, the use of analytical models decreased significantly. Kozan & Higgins [8] developed an analytical model to estimate delays for individual trains and track links in an Australian rail network. They compared the results to that of obtained from a simulation algorithm. For 93% of the 157 scheduled trains the analytical model's delay estimates were within 20% of that of the simulation algorithm's estimates which was deemed more accurate. This inaccuracy was attributed to the sensitivity of the analytical model to slight differences in the assessed and actual delay distributions. This paper therefore highlighted one of the short-comings of analytical models when compared to simulation models.

When it however comes to optimising train schedules, heuristic algorithms are used such as Job Shop, genetic and Tabu-search algorithms. These will be discussed in later sections.

3.2.1 Queueing models

Queueing theory, originally referred to as telegraphic theory, has been developed since the 1920's for telecommunication services. The application of this theory has since expanded to the computer, manufacturing, retail services and transport industries.

Queueing processes are usually described by six characteristics listed by Gross, et al. [9] as:

1. Arrival pattern of customers
2. Service pattern of servers
3. Number of service stages
4. Number of service channels
5. Queueing discipline
6. Capacity of the system

The arrival pattern in most queuing models is stochastic of nature and follows a certain probability distribution of inter-arrival times. It can however also be deterministic depending on the systems being modelled. When setting up the parameters for arrival it is necessary to know if agents can arrive in bulk (i.e. simultaneously), and if so, the probability distribution of the size of the bulk. In some models an agent can decide not to join the queue upon arrival - this is referred to as *balked*. In some cases an agent can enter a queue and then after a while lose patience and leave the queue (it is referred to as *reneged*). Another case may be when there is more than one queue and an agent switches from one queue to another. This is called *jockeying*. Further on, when an arrival distribution does not change over time it is referred to as *stationary*, and when it does change, *nonstationary* [9]. Note that in rail systems *jockeyed* and *reneged* arrivals are not considered. Trains cannot practically arrive in bulk because of headway constraints forcing trains to have a certain time or distance buffer between them. Headway constraints are enforced for safety purposes, and are applicable in any railway system. Similarly trains cannot renege or jockey a queue (waiting track) if the driver becomes impatient. It is however possible for a train to *balk* (note that there is a difference in meaning between *bulk* and *balk*). When a serious disruption occurs on a route, following trains can be rerouted where possible, or even be cancelled.

Similar to arrival patterns, service patterns also have distributions describing the time an agent spends being serviced. Agents can also be serviced in bulk or singularly. The service time however can be influenced by the size of the queue or arrival pattern. In such a case it is referred to as a *state-dependent* service, but generally arrival and service patterns are assumed independent [9]. Another aspect of service time, as with arrival patterns is that it may change over time. For example, when learning takes place and the service process becomes quicker and more efficient. The same terms as previously mentioned - *stationary* and *nonstationary* – are used for such service processes. This is not usually applicable in rail systems, as trains have specified dwell times at stations. In South Africa, however passenger train drivers may compromise specified dwell times, to either catch up lost time because of a delay or dwell longer because of passenger over-crowding.

The manner by which an agent is chosen for service from a queue is referred to as queueing discipline. The most common discipline is the first-come-first-served (FCFS) principle, and in some inventory systems last-come-first-served (LCFS) principle applies. Other systems have priority schemes which are usually either called *pre-emptive* or *non-pre-emptive*. *Pre-emptive* priority is when a high priority agent enters a queue, service on a low priority agent will be paused and the high priority agent will be serviced first. In the case of *non-pre-emptive* priority the high priority agent will be moved to the front of the queue but will only be serviced when the agent being served at that moment is finished. Passenger rail systems mostly work on the FCFS principle, whereas freight rail systems might have different disciplines which take into account the importance of the freight content [9].

Some systems have limited queues which create a limited system capacity, such as a doctor's waiting room with a number of chairs. However some queueing systems have infinite length, as in the case of

judicial processes or waiting lists. In the case of rail systems where stations and sections are the servers, queues are limited.

Queueing systems can have more than one service channel. In general it is preferred to have a single queue feeding multiple channels e.g. customs at airports and railway stations with more than one platform. This usually applies for systems where the agents have no preference as to which service channel they want to use. In for example a bank, where different services are offered at each queue customers will line up in multiple queues [9].

The last aspect of queueing systems is stages of service. Systems may have more than one service stage and manufacturing systems are good examples of this. Parts will for instance be assembled, and then be moved forward to be checked for quality. If the quality is not satisfactory, the assembly will be fed back to the previous stage or otherwise the assembly will move forward to the next stage [9]. Passenger rail systems only have one service stage, while freight trains may have more (i.e. freight being unloaded and then the train moves to the hump yard etc.)

The following points summarize queueing systems:

1. An agent arrives according to a certain probability distribution or fixed inter-arrival time.
2. The agent then enters or does not enter the queue, depending on the type of system.
3. The agent then moves from the queue to get serviced for a duration specified by the modeller. This can be for a random or fixed time period.
4. After the agent is serviced it leaves the system and the next agent in the queue is serviced, depending on the queueing discipline.

Huisman et al. [10] developed a queueing network model to compute the long term performance of rail networks. To achieve this, a decomposition of the network in its detailed components was necessary. These components include stations, junctions and sections. The network performance was measured by the mean delay and delay probability of the trains arriving at their destinations. Because train movements are not known over the long term, assumptions were made to simplify the modelling of stations. One of the assumptions is to model the halting tracks outside of the model. Thus when a train finishes its route it exits the model and is stored in a queue outside the model. The halting track is where the train starts its route, and where the passengers alight or board the train. The next train can only enter the model after the train on the halting track has departed.

The occupation times at the halting tracks are assumed to be exponentially distributed and equal for all train types. The stations are modelled as multi-server (since stations have more than one platform) queueing systems with Poisson arrival distributions.

The same principles were applied to junctions and signal blocks, except that these were single server queues. If a junction is occupied, the next train falls into the queue, until the junction is clear. This occupation time is also exponentially distributed.

Track sections between stations were broken up in signal blocks, with each block acting as a separate queuing system. Bottlenecks and delays were then calculated by adding up all the waiting times in the queues. These waiting times were compared to the real delay durations of the trains.

The model showed good accuracy even though a probability distribution was used for inter-arrival times to determine arrivals for trains, instead of using a timetable. Yuan & Hansen, [11] and Meester & Muns [12] both emphasise the lack of queueing models to consider timetables, since it is reliant on probability distributions for inter-arrival times. Moreover fixed arrival and departure times were also not considered and the impact of speed variations with different train types was neglected. Huisman et al. [10] however suggested a way to capture speed variances among different train types by ignoring block (signalling) sections in a section between stations. The model does however include one block section before and after each station, to insure that trains do not arrive in bulk at stations. Furthermore the number of trains allowed in a section was limited to the number that would have been allowed in the case with signal blocks. Speed variations was accounted for by for instance, if a section has five signalling blocks the middle three sections will be removed from the model and only the first and last sections will be included. This allows enough distance for a train with a different speed to have a significant variance in free running time. (Here free running time refers to the time a trains travels between stations without any disruptions). Huisman's model was applied to two major lines of the Dutch network namely. Rotterdam – Utrecht and Den Haag – Utrecht. The traffic on this network is extremely heterogeneous with three different train types (implying three different train speeds) running three different services.

de Kort et al. [13] also applies a similar queueing model based on Wakob's Approach, on Den Hague station in the Netherlands. Wakob's approach breaks up all the components of a station and analyses them independently as separate queues. Arrival and service times are both assumed to fit an Erlang distributions resulting in $E_k(\lambda)/E_l(\mu)/I$ queues for the whole queueing system. de Kort et al. [13] argues that service time variations should be dependent on running time and dwell time variations, instead of independent probability distributions. It was found that this approach over-estimates delays or alternatively models the "worst case scenario". This may be related to the fact that Wakob's approach returns the upper bound of the delay duration instead of the mean and standard deviation. This approach is thus inappropriate for delay propagation analysis, however it can be useful for capacity planning purposes [13].

Queueing models can serve as a good alternative to simulation to estimate delays, however as previously mentioned modelling large networks becomes difficult to solve analytically. Kozan & Higgins [8] explains this complexity of train networks with the following:

"A train network is complex in that it includes many intersections, uni- and bidirectional track links of various lengths, sidings, and track capacity. Train services vary with different upper velocities, slack time, scheduled stops, non-uniform departure times, and include train connections as described in the

introduction of the paper. In the case of train connections and intersections, a train can suffer a delay from another that is scheduled much earlier and from a different part of the network.”

“As well, the distribution of arrival times for each train at any station or intersection depends on the distribution of current delay, which can be different for each train service. Hence, delay to both the trains and at stations (or intersections) are interdependent. Therefore, the calculation of expected delay requires a solution of equations.”

3.2.2 Job shop models

Branch and bound algorithms have been used to develop and optimize timetables. These models transform train networks into large job shop models. Typically trains will be jobs and stations and sections will be machines. In job shop models there are a number of different jobs that need to be completed by a number of machines. A job will have a specified time and order it has to spend at each machine. For example *Job A* will use *Machine 1* for 2min, then *Machine 2* for 5min and lastly *Machine 3* for 3min. *Job B* will first use *Machine 2* for 3min then *Machine 1* for 5min and then end off with *Machine 3* for 1min. Figure 3-1 shows an illustration of this simple model. It is important to note that each machine can only work on one job at a time. This means that when *Job B* is finished with *Machine 2*, *Job A* can move to *Machine 2*. Similarly when *Job A* is finished with *Machine 1*, *Job B* can move to *Machine 1*. Whichever *Job* first finishes using *Machines 1* and *2* then moves to *Machine 3*. The other *Job* will then have to wait for the first *Job* to finish before moving to *Machine 3*. In the example illustrated in Figure 3-1 both *Jobs* will arrive at *Machine 3* at the same time. In such cases priority rules can be implemented. Nevertheless problems of this nature, create the need to determine what the optimal sequence of machine usage is, i.e. which job should utilize which machine, when? Branch and bound algorithms are used to solve these problems. For further explanations on Job shop models and branch and bound algorithms refer to [14].

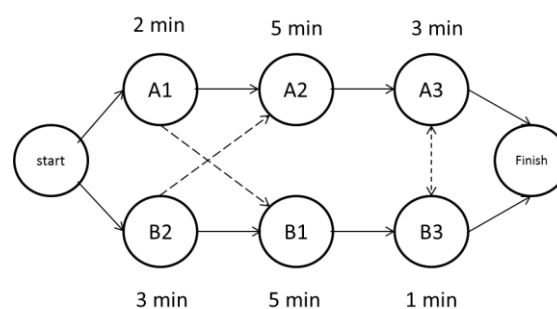


Figure 3-1: Simple job shop model

Rail networks can be similarly modelled, seeing trains as *jobs* and stations, sections and junctions as *machines*. There does however exist key differences between train network models and classical job shop models [14]. These are listed as follow:

- Jobs and machines do not have lengths as do trains and sections. While moving from one section to another a train’s “head” will occupy the next section while the “tail” will occupy the

current section. A train may thus occupy two sections at a time, whereas jobs can normally only occupy one machine.

- Train acceleration, deceleration and cruising speed for a specific section cannot always be pre-defined, since it is dependent on the train in front.
- Trains can visit sections more than once, whereas jobs are mostly assumed to visit machines only once.
- Passing facilities such as passing loops on rail sections are equivalent to capacitated buffers or parallel machines. These are very difficult features to model with a standard job shop model.

In the paper of Burdett & Kozan [14] it is explained how these differences were incorporated in order to produce realistic results.

D'Ariano et al. [15] also developed a job shop model for the Dutch railway network. Figure 3-2 shows a small network on which the model in Figure 3-3 is based. Note that each block section is represented by a *machine* or a *resource*, as referred to in this paper, and *Trains A and B* are the jobs. A minimum headway of one signal block between trains is modelled and indicated by the dotted arrows in Figure 3-3. For example, *Train A* can only enter *block 5* when *Train B* has exited *block 7*.

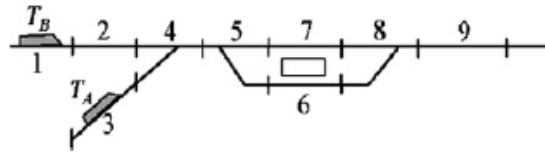


Figure 3-2: Small network with 9 block sections and two trains [15]

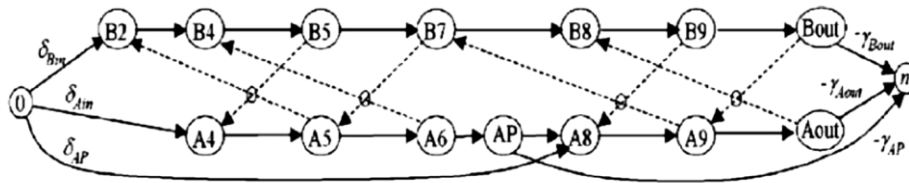


Figure 3-3: Job shop graph of two trains [15]

This model was expanded to model the Schiphol rail network which includes the stations of Nieuw-Vennep, Hoofddorp, Amsterdam Lelylaan and Amsterdam Zuid. The network consists of 86 block sections, 16 platforms, traffic in two directions and 54 trains.

The model wished to solve the train scheduling problem for real-time rail network management. The objective function is to minimize the maximum secondary delays at all stations by all trains. It was found that these algorithms perform better than the despatching rules commonly used with regard to average and maximum delays.

Burdett & Kozan [9] used a hybrid job shop model with time window constraints to solve the train scheduling problem when adding additional train services. In their later work [7] they again used the job shop approach but then further refined the solution using simulated annealing and local search meta-heuristics. This allowed them to shift trains more easily and feasibly within the solution plane.

3.2.3 Tabu search

Tabu search is a meta-heuristic algorithm which memorizes the most recent local optimum. As soon as a solution is found which is better than the previous best solution, the algorithm will store it and discard the previous best solution (i.e. the solution becomes tabu). This also implies that the algorithm will never return to the same solution twice. The Tabu search thus eliminates the possibility for the search to get stuck on a local maximum and continually searches for new local optima in the solution space.

Corman et al. [16] compare a Tabu search algorithm to a local search algorithm and various hybrid algorithms previously developed [15], [17] to solve routing and scheduling problems in the Dutch rail network. The study focussed on a bottleneck at the dispatching area of Utrecht Den Bosch, which consists of 191 block sections, 21 platforms and 50 km of track. The algorithms had to search out of 356 possible routes for the best solution. The results showed that the Tabu search algorithm reached better solutions faster, compared to the other algorithms.

Similar conclusions regarding to quality and speed of solutions reached by Tabu search methods were found by Higgins et al. [18] which solved the problem of a single track line with occasional sidings for opposing trains to pass each other.

3.2.4 Genetic Algorithm

Genetic algorithms are very effective and robust algorithms to determine global optima. Gradient based methods, such as Steepest Accent, Conjugate Gradient or Lagrangian Multiplier, usually converge faster to local optima or a local optimum than a genetic algorithm, however in cases of multi modal functions they may miss the global optimum more often than not. Genetic Algorithms are based on the theory of genetic evolution where the fittest genes in a chromosome survive and the weakest genes die away in the process of reproduction. To put in differently, the offspring of two parent chromosomes will only consist of the best genes found in both parents. In this way continual improvement in fitness takes place with every generation [19].

Considering the algorithm, each solution is represented by a chromosome. Stochastic mutation of some of these offspring is brought in at pre-determined instances in order to make sure the algorithm does not get stuck on a local optimum. The numerical values of a solution's parameters are converted to a series of binary digits, and each parameter is then represented by a gene. When a gene thus evolves the digits of its binary code change to either 1 or 0 [19].

Genetic algorithms are not commonly used for solving train scheduling problems, however Higgins et al. [18] used a genetic algorithm to solve a single line train scheduling problem. In this study each

gene contained three attributes, namely: the delayed train, the train with the highest priority or right of way and the track section where the conflict will occur. With each parent then consisting of six genes (e.g. six train schedule solution), the fittest two parents are chosen to mate and produce two children with genes from both parents with a single randomly selected crossover point. The genes before the crossover point are transferred the first child, whilst the genes after the crossover point are transferred to the second child. Mutation in this algorithm has a very low probability, however when mutation happens and the conflict gene changes, and the neighbouring genes also change. Changing only one conflict gene by mutation is not good in train scheduling problems [18]. The genetic algorithm in this study proved to outperform the Tabu search and local search heuristics which the authors also used to solve the same problem.

It seems that most of the cases where genetic algorithms were used, were in cases of single track lines with traffic in both directions [3] [20] [21].

3.3 Simulation models

Saayman & Bekker [16] explains simulation as an attempt to solve real world problems, by first building a model that represents the current state and operation of a system as realistically as possible. This is achieved by making argued simplifications and assumptions. The model can then be used to solve, experiment or optimize the modelled system. Saayman & Bekker [16] goes further to explain that simulation allows the modeller to include the stochastic nature of real world systems. It allows for big scopes and high complexity systems. It is however difficult to validate a model, since the whole point of simulation is to forecast the effects of change to a system before spending capital to implement the intended change. Model validation is usually done by comparing the “current state” model to actual system behaviour. In this way the modeller can make the assumption that the model is a realistic representation of the system. Simulation is thus a tool that should be applied with care, since getting answers is easy but getting realistic answers is a fine skill [16].

In the railway environment there exist two types of simulation approaches to modelling train operations, i.e. macroscopic and microscopic. This Section will explain the difference between the two approaches and give examples of where they were applied.

3.3.1 Macroscopic simulation models and software

Macroscopic models are used to evaluate the operation of a transportation system, and uses statistical data to describe trains’ behaviour. Detailed individual train behaviour and movement are thus simplified in order to be able to model larger systems [22].

NEMO (Network Evaluation MOdel), a macroscopic model developed by IVE and the Austrian Federal Railways, is used for strategic planning and evaluation of infrastructure and operational complexities. Radtke and Hauptmann [23] used NEMO to model large parts of the German railways macroscopically and then combined the outputs with a microscopic model built in Railsys. Microscopic models will be discussed in Section 3.3.2, however microscopic models relate to the

more detailed approach to modelling trains. Radtke and Hauptmann's [23] approach therefore suggested ways to combine macro- and microscopic railway models, even though the results and computational performance was not compared to similar approaches such as developed by Schlechte et al [24].

Hwang & Liu [25] developed a macroscopic simulation model to forecast the effect of increasing demand for railway capacity of the Taiwan regional railway system. The objective was not only to model increase in the line capacities but to also improve the efficiency of the current capacity. The model's objective was the accurate estimation of knock-on delays (secondary delays), as a result of a primary delay. The following input parameters were used to represent the network:

- Railway condition – the line, stations and track layouts of the stations
- Traffic condition – minimum dwell time and scheduled timetable
- Control condition – minimum headways, section capacity and recovery time

With these parameters the model was run assuming no delays, i.e. strictly following the scheduled timetable. To determine the effect of a primary delay on the network then, a delay event had to be created. This event or primary delay is defined by four parameters namely, location of delay, delay start time, delay release time and the magnitude of the delay. The magnitude of the delay is simply the difference between the delay start time and delay release time. The resulting secondary delays were thus one of the outputs of the model. These delays were then used to create a simulated timetable.

To validate the model, actual train operating data was used. The arrival-departure time data of a specific day was retrieved from the Centralized Train Control database of the Taiwan Railways Administration. Actual delay data was also collected in order to compare with the simulation output. A route conflict delay was chosen as the real event that serves as the primary delay. The model proved to be within 120 seconds of the actual delay time 77.5% of the time, with 62.5% of the time within 60 seconds. Figure 3-4 shows the Marvey diagram² of the normal timetable without any delays and Figure 3-5 shows the diagram for the simulated timetable. It is clear that a delay occurred between Shongshan and Taipei stations and the next seven trains were affected by it. Hwang and Liu [25] went further to compare different delay reduction strategies and how they influence the total secondary delays. The effects of three strategies are shown in Figure 3-7. It is interesting to note the exponential relationship between primary (or first delay) and secondary delays (or knock-on delays). This can be explained by the fact that the larger the primary delay, the harder it is for a train to recover any of the lost time. A train is naturally limited in ability to use these three strategies to recover the lost time created by the primary delay. A train has a minimum allowed dwell time at stations and is also

² A Marvey diagram is a time-distance diagram of a train's journey from its origin to destination station. Lines running at a positive inclination are generally referred to as "up-trains" or trains moving in the "up" direction and lines at a negative inclination are referred to "down-trains" or trains moving in the "down" direction. The lines in the two different directions can only cross if the trains are running on a double line section, or if there is a passing loop. Furthermore, the steeper the lines the faster the train is running. Lines running horizontally therefore indicate a train standing still.

subjected to speed limits on sections. These limitations thus translate into knock-on effects on later trains which result in an exponential growth in the total delays.

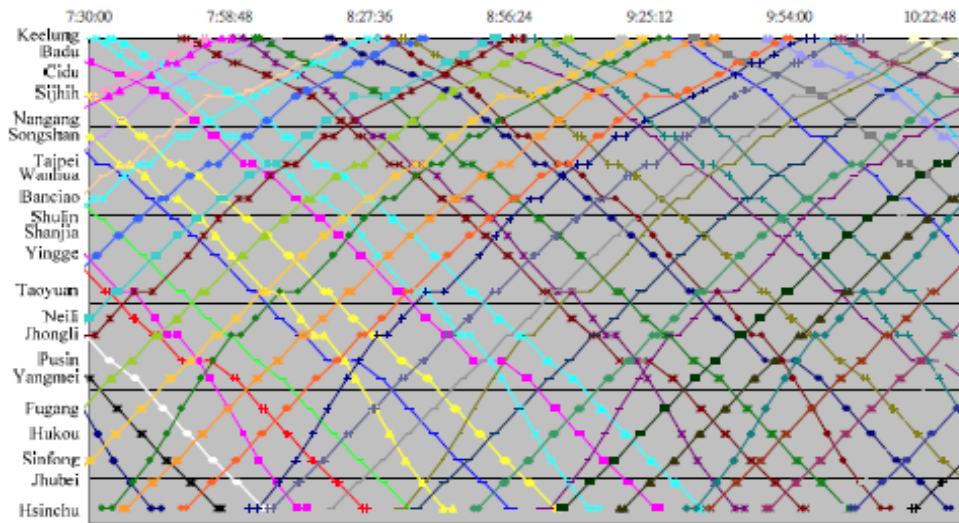


Figure 3-4: Normal timetable without delays [25]

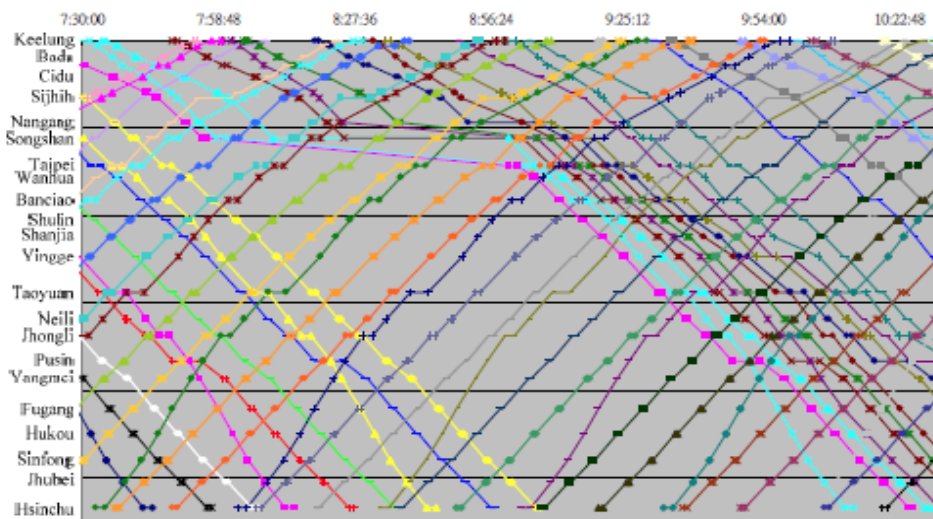


Figure 3-5: Simulated timetable diagram with delays [25]

Middelkoop & Bouwman [26] demonstrated the use of *Simone* simulation software to model the entire Dutch rail network. The software requires the following as inputs to the model:

- Infrastructure data
- Timetable
- Simulation specific parameters
- Network properties with regard to disruptions and disturbances
- Operational rules
- Statistical indicators for the simulation output

The software then produces the indicators pre-specified by the user and an animation of the network operation. Figure 3-6 shows an example of the animation output *Simone* produces. The figure shows a part of the Dutch rail network and all the trains operating on it. Each type of train has a unique colour. Most parts of the model were constructed by the software's automatic model generator. The model included 600 stations, 1100 track sections and 350 trains which is significantly large. The model was able to show (see Figure 3-7) for example the punctuality of trains in certain parts of the network and the relationship between initial delays and sum of delays (as done by Hwang & Liu [25]).

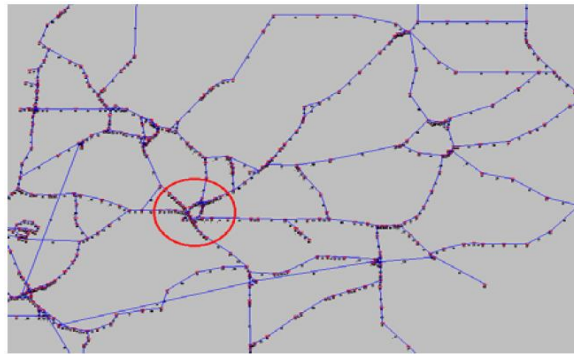


Figure 3-6: *Simone* simulation animation output [26]

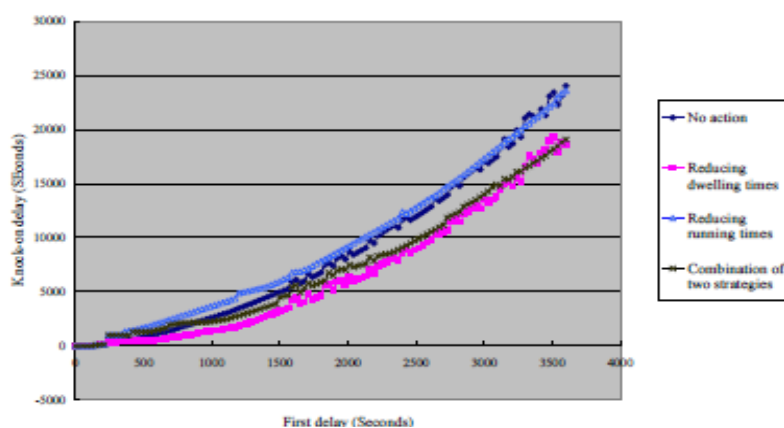


Figure 3-7: Total knock-on delays at the destination station [25]

On the East coast United States of America, one of the major railway companies, CSX, used Anylogic simulation software to create a visual emulator to replay past system behaviour of their entire network on a GIS map to better understand density, flow and congestion processes. Train movement data was imported to Anylogic from databases to pre-define train behaviour. It should be noted that this model was built without the use of Anylogic's Rail Yard Library [27].

CSX also used Anylogic's agent based modelling ability to model a supply chain network to determine the ability of the operator to fulfil the demand of coal trains and the ability to stage empty trains. In this model the trains were modelled as agents moving across the network and making decisions based on built-in intelligence.

3.3.2 Microscopic simulation models and software

Microscopic simulation models are used to model actual operations of trains. Detailed train movements and behaviour are considered together with relevant infrastructure details. These models are used on smaller networks and can in some cases be optimized. In general, the objective of microscopic models is to test schedules and simulate the effect of disruptions [22].

Train networks can be simulated in two ways. One is time-based modelling where a time span is broken up into equal intervals and train movement is calculated at each interval. This is a very realistic representation of train movement; however it requires a large amount of information with every update, making it computationally intensive. Time based models are typically used in signalling layout design and energy consumption analysis [28].

The second way of simulating train movement is event-based. This method is similar to queueing models discussed in Section 3.2.1. The train's movement is described in terms of a chain of events. For example, the train arrives at a station at a specified arrival rate and dwells for certain time period. The train then leaves and enters a track section which marks the start of the next event. Each event's duration is characterized by a certain probability distribution. Event-based models may reduce computational time significantly compared to time based models, however train movement updates are not synchronised between events [28].

van Dijk [29] suggested that queueing theory and simulation can be combined. He argued that the advantages of queueing theory (generic components, few detailed data needed) reduces the disadvantages of simulation namely, high level of complexity and detailed data required. In the same way simulation's advantages (i.e. real-life complexity and real-life uncertainties) reduce queueing theory's disadvantages namely, over-simplification and unrealistic constraints.

Azadeh et al. [30] used a Visual SLAM coding language to develop a simulation model of a complex rail system consisting of 50 stations and both passenger and freight trains. An analytical hierarchy process (AHP) method was used to weigh the qualitative and quantitative inputs and outputs which were then converted to a data envelopment analysis (DEA). The objective of the model was to find ways to increase passenger train reliability and decrease turn-around time of both passenger and freight trains.

Ho et al. [28] developed a general-purpose multi-train simulator which enables the user to model without carrying out program code modifications. The simulator has been used in Hong Kong and China for studies of traffic control at conflict areas, scheduling optimization and energy management of trains.

In the railway literature it was experienced by the author that the majority of simulation work is done to determine capacities of complex single line networks. Single line routes with different loop lengths and train types (e.g. passenger and freight trains) are extremely complex to schedule and to calculate capacity. Simulation software packages such as RTC (Rail Traffic Controller) and OpenTrack are then

used to simulate these complex networks. OpenTrack, which is similar to RTC, uses microscopic models to simulate rail operations and are based on user defined train, timetable and infrastructure databases [22]. OpenTrack was developed by the Swiss Federal Institute of Technology's Institute for Transportation Planning and Systems and is mainly used to test and evaluate operating schedules and timetables[22].

Schlechte, et al. [24] used OpenTrack to simulate the Simplon corridor between Switzerland and Italy. The objective was to maximise a utility, which can be a constant or monetary value, of the allocated trains. A LP optimization algorithm was written in CPLEX to determine the optimal schedule. This schedule was then simulated in OpenTrack to test its validity. The authors claim that their method allows for solutions that are comparable in quality to most sophisticated manual schedules, but were produced much faster than manual schedules.

RTC from Berkeley Simulation Software, was developed in the United States of America and is very popular for simulating single line freight and passenger rail networks. Dinger et al. [31] used RTC to study the impacts of the various aspects of train type heterogeneity on the planning of rail operations of North American railroads. Furthermore suggestions were made as to how delays, caused by heterogeneity, can be reduced by certain operating strategies.

Mei-Cheng Shih [32] also applied RTC to a North American rail network that was originally designed for low traffic density (i.e. infrequent passing loops) and short trains. An increase in demand of commodity flows thus created the need for an effective capacity expansion strategy. RTC was used to simulate experiments for various expansion alternatives.

Similar work was done by Sogin et al. [33][34] where trains on North American networks were mixed in terms of speeds and priorities. Passenger trains running at speeds up to 110 mph and enjoying highest priority were sharing the same infrastructure as freight trains running at much lower speeds. This heterogeneity and increase in traffic required simulation to determine operational and expansion strategies.

The author's own experience in literature of both OpenTrack and RTC has shown that OpenTrack relies on scheduling data as input. Detailed itineraries have to be specified by the user before the simulation is run. If an invalid schedule is used the software will produce an error. RTC however has the ability to calculate a valid schedule given the starting times of each train. RTC thus uses sophisticated passing logic to determine valid schedules for rail networks. For this reason it is used for capacity modelling of highly heterogeneous rail systems. RTC has been applied very successfully in several commercial rail projects around South Africa.

Unlike specialised rail simulation software such as RTC and OpenTrack, Anylogic (which is a multi-method general simulation software package) has also been used for microscopic rail simulation models. Anylogic was used to determine the capacity of a rail maintenance yard for Australia's largest rail freight operator. The model was built with the help of Anylogic's Rail Yard Library. The model

was able to test amongst other, train configurations, track utilisation, scheduling of activities and train movements within the yard [35].

3.4 Train system punctuality and reliability

The purpose of modelling train networks is primarily to estimate punctuality, since this and a safe journey are the key factors to passenger satisfaction [36]. Gylee [37] also makes this point by explaining punctuality as “the ability to achieve a safe arrival at a destination to an advertised timetable”. Punctuality will thus be discussed in this Section.

A major aspect that influences punctuality is reliability. Rieveld et al. [38] uses reliability to define punctuality and refers to it as the same thing. However in this study reliability will refer to the ability of the system to not fail to such an extent that a delay will be caused. The system here includes the electrical infrastructure, perway infrastructure and rolling stock. Each of these components has their own way of defining and quantifying reliability. For example a recent study by Conradie [39] quantified the reliability for rolling stock of the Passenger Rail Agency of South Africa. However a failure in that study did not necessarily mean a failure causing a delay. A failure in this study will mean a failure that caused a train to either stop or be slowed down enough to cause a delay.

Punctuality is not always measured in the same way. The most common way calculating it is by the percentage of trains that are punctual. Norway determines a train’s punctuality at its destination station [40], even though some argue that punctuality should be determined at each stopping point along a train’s route [41]. Olsson [40] did a correlation analysis on the following factors that may affect punctuality:

- Temporary speed restrictions
- Construction work
- Infrastructure capacity utilisation
- Occupancy ratio
- Number of passengers
- Departure and arrival punctuality
- Operational priority rules

High capacity utilization is widely assumed as a large contributor to secondary delays, however it was found that the number of passengers were more correlated to poor punctuality than capacity utilization (capacity utilization refers to the ratio of trains running on a line to the maximum number of trains able to run on the line). Further on operational rules at crossings and the management of boarding and alighting passengers (which influences departure punctuality) seemed to be the most significant factors pertaining to overall train punctuality.

3.5 Summary

This chapter discussed the various ways to model and schedule train networks. Firstly pure analytical models were covered which showed that networks can be modelled accurately without advanced computational methods. They are however very limited in terms of scope and network complexity.

Secondly heuristic methods were discussed. It can be concluded that these methods are very effective in optimising large complex networks. It allows the modeller to find global optima amid a solution plane consisting of many local optima. Optimising train schedules for dense rail networks seems to be possible with the right combination of these heuristic algorithms.

Lastly the use of simulation was discussed. Simulation allows for very large scopes and even entire networks to be modelled [26]. It also has the ability to include important infrastructure detail and also simulate reality fairly accurately. Moreover it possesses the ability to animate the model making the complex nature of a rail network visual and easier to understand. In the rail environment train operation simulation is differentiated between micro- and macroscopic simulation models. Macroscopic models often include entire transport systems of which rail can be part of, or in other cases very large rail networks. Microscopic models are much more focussed on the detail aspects that influence train movement and are used for smaller networks. Simulation models have the ability to calculate the capacity and expansion strategies of complex rail networks.

4. Model development

In this chapter a model will be developed to model heterogeneous rail traffic operating in an aged system. It will thus model rolling stock and overall system reliability and how it affects train punctuality. Since the model uses statistical methods to describe reliability and simplifies certain aspects that determine train movement, according to Nash [22], this model can be classified as a Macroscopic model.

In Section 4.1 the basic outline of the model will be explained. The infrastructure sub-model will be explained in Section 4.2 while the two fundamental components of the model will then be explained in Sections 4.3 and 4.4. The limitations and assumptions of the model will then be explained in Section 4.5, and finally the chapter will be rounded off with the concluding Section 4.6.

4.1 Basic outline

In this Section the inputs and the outputs will be discussed to give an overview of the model. The model itself will be discussed in the later sections.

4.1.1 Outputs

To understand the objective of the model one has to first look at the outputs of the model. The objective of the model is to give a measure of punctuality of an unreliable rail system. The idea of the model is to see the effect on a system when *change* is implemented. Thus in this model the *change* will be introducing heterogeneous train traffic into a rail system which was always operated with homogeneous train traffic and to increase the reliability of the system as a whole. To draw this back to the objective of the model - the punctuality of the trains will be measured with the change in heterogeneity of the traffic (i.e. changing the mix of train types) and change in system reliability.

To measure punctuality one has to first look at what the criteria are for a train to be punctual. The most common measure used in Britain and Europe is 5 minutes from a train's scheduled arrival time; e.g. when a commuter train arrives at its destination station within 5 minutes of the scheduled time it is regarded as punctual [42]. Any time a train arrives later it will be regarded as late and delayed. This will also be the measure used in this model.

The outputs for the model are thus:

- Total minutes delay over a given operational period
- Number of delays
- Average delay duration

The key parameter is Total minutes delayed since it is the parameter by which train punctuality is measured. Subsequently, since Total minutes delayed is a function of Number of delays and Average delay duration, the latter two parameters will also be measured to understand how they are influenced when *change* is introduced in the system.

4.1.2 Inputs

In order to produce the outputs mentioned in Section 4.1.1 accurately, one has to input the relevant constraints and variables that describe the system and the environment in which it operates.

To describe the environment in which this model operates, the system will be broken up into five parts: stations, perway, signals, rolling stock and over-head traction equipment. The case study discussed in later chapters, includes trains that can regenerate electricity back into the grid when it either brakes or runs free. In reality this function may have an influence on how trains are scheduled because of various electrical infrastructure constraints such as the inability of the current Western Cape rail infrastructure to store the regenerated electricity. In effect this means that when a train generates electricity, there needs to be another train that consumes that electricity. This function will however be ignored in this study. The inputs that describe the environment of the model are the following:

- Stations
 - GPS coordinates – indicate locations of the stations on the route.
 - Maximum capacity. The number of platforms or tracks in a specific direction in a station.
 - Dwell times. The time a train dwells in a station before departure. This value will be fixed since variation in dwell times for passenger services are usually a matter of seconds while delays are measured in minutes. Randomness in dwell times will therefore not have a significant influence on the outputs.
- Perway
 - GPS coordinates. When modelling a train system or network, the geographical details of the route or piece of network being modelled, is an essential input. Some models don't necessarily use geographical information such as GPS coordinates, but they still ensure that the dimensions of the route or network are correct. This model aims to model directly on a GIS map.
- Signals
 - GPS coordinates. The exact location of each signal along the modelled route is required.
 - Operational rules. This includes min headways at stations, sections and junctions measured in signal blocks.
- Rolling stock
 - Speed and acceleration properties. The defining factor that classifies this model as a heterogeneous train traffic model is the difference in speed and acceleration properties of the two different types of trains (old and new trains). Reliability also differentiates between the train types, and will be further explained in later Sections.

- Schedule. The schedule will be used to determine when a train must depart from its starting station and trip times to its end station. It will however be revised with each scenario to accommodate the new trains.

These inputs contain no randomness, as they are only there to describe the modelled environment. With only these inputs the trains will run perfectly according to schedule, departing on time and arriving at their destinations on time.

Finally, the stochastic inputs are in the form of delays, or otherwise referred to in literature as disruptions. Trains experience either a primary delay or a consequential delay. Goverde [43] explains the definition of a primary delay and a secondary delay (consequential delay) as follows: “A *primary delay* is the deviation from a scheduled process time caused by disruption within the process” and “A *consequential delay* is the deviation of a scheduled process time caused by conflicting train paths or waiting for delayed trains”. The input to the model will be a primary delay, and the delays caused on the following trains as a consequence will then be referred to as the consequential delays. A primary delay is mostly caused by a failure of either the train or track infrastructure. Delays can also be caused by passenger overcrowding or other passenger related accidents. Regardless, a primary delay has the following three parameters:

- Location of train where primary delay occurs
- Time when primary delay occurs
- Duration of the primary delay

The location of the primary delay will depend on the nature of the system being modelled. If primary delays tend to occur more in a certain region of the network, the probability of a primary delay occurring there should be adjusted accordingly. However, the location of a primary delay will have no influence on the sum of the consequential delays if only one line is modelled. This means that regardless of where a train is being delayed on a single line, the consequential delays will add up to the same amount. This is true for both homogeneous and heterogeneous traffic. This principle is however not valid when modelling a route with converging or diverging lines.

In Figure 4-1, a simple Marvey diagram is shown to illustrate the movement of 5 trains running through 5 stations all at the same speed and headways. Train 1 then gets delayed for 23 minutes at Signal 2 which causes the following trains to also be delayed. The following trains will all have to wait until there is a minimum required headway of 5 minutes between it and the train in front, before continuing to the next station. The consequential delays are thus calculated as follow:

$$D_{i+1} = D_i - H_{i+1} + h \quad i = 0,1,2, \dots, n \quad D_i > H_{i+1} - h \quad H_{i+1} > h \quad (1)$$

And the sum of delays:

$$D_{total} = \sum_{i=0}^n D_i - H_{i+1} + h \quad (2)$$

Where: D_0 = *Primary delay*, $D_{1,...,n}$ = *Consequential delay*, h = *minimum required headway*, H = *headway*. The sum of delays for this case is then 1:05:00. If one refers to Figure 4-2, where Train 1 was delayed at Signal 7 for the same duration, the sum of delays is also 1:05:00, showing that for homogeneous traffic the location of the primary delay has no influence of the sum of delays.

A similar calculation was done on a case with heterogeneous traffic where Train 3 was given faster speed properties resulting in a 3 minute shorter trip time. This implied that different headways at the start station had to be determined in order for the all the trains to arrive at the end station at equal intervals. Again a primary delay of 23 minutes was initiated at Signal 2 for the one case and Signal 7 for the other (see Figure 4-3). It is important to note that after Train 3 is delayed, it continues its trip at the same pace as the slower trains, since it is not possible for trains to pass. The sum of delays was found to be exactly the same as in the case of homogeneous traffic.

Thus for either homogeneous or heterogeneous train traffic on a single line, the location of a primary delay will make no difference to sum of delays. It must be mentioned that this does not apply to networks that consist of different lines converging and diverging at junctions, but for the purposes of this model the rule applies.

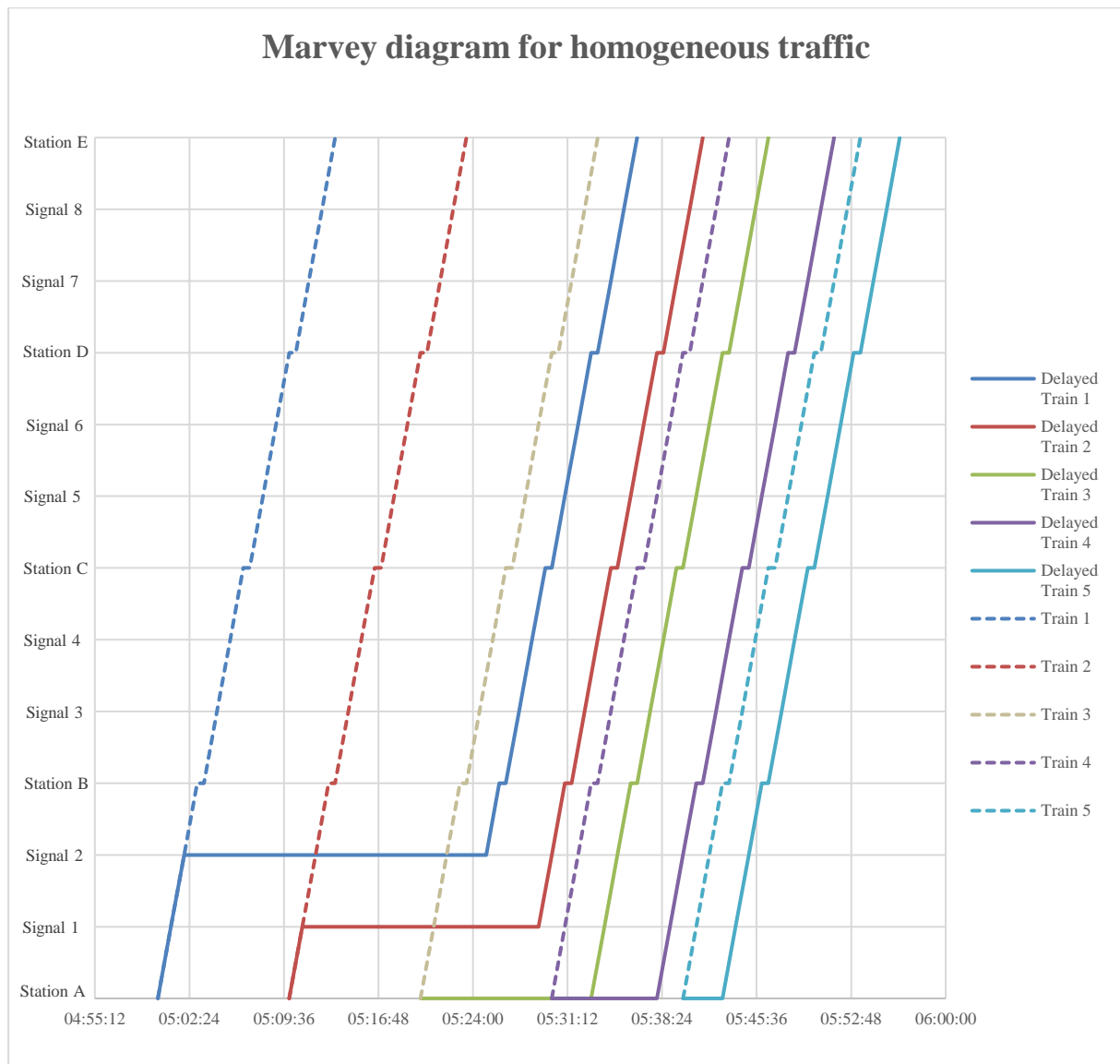


Figure 4-1: Marvey diagram for normal and delayed homogeneous rail traffic consisting of 5 trains with the primary delay occurring at Signal 2.

Now moving on to the next parameter, namely the time when a delay occurs. This is not determined stochastically. It may be difficult to determine a distribution of when delays occur from available data, since most delays are caused by train or infrastructure failures which are not necessarily dependent on the time of day. Some delays however are not necessarily caused by failures but rather by human disruption. Train overcrowding and train driver slackness are examples of such delays. It is also important to know that failures may occur in either rolling stock or infrastructure without causing a delay. For instance sub-components on a train may fail, such as a traction motor, but it won't cause the train to stop or even lose speed as the other motors are adequate to keep the train running.

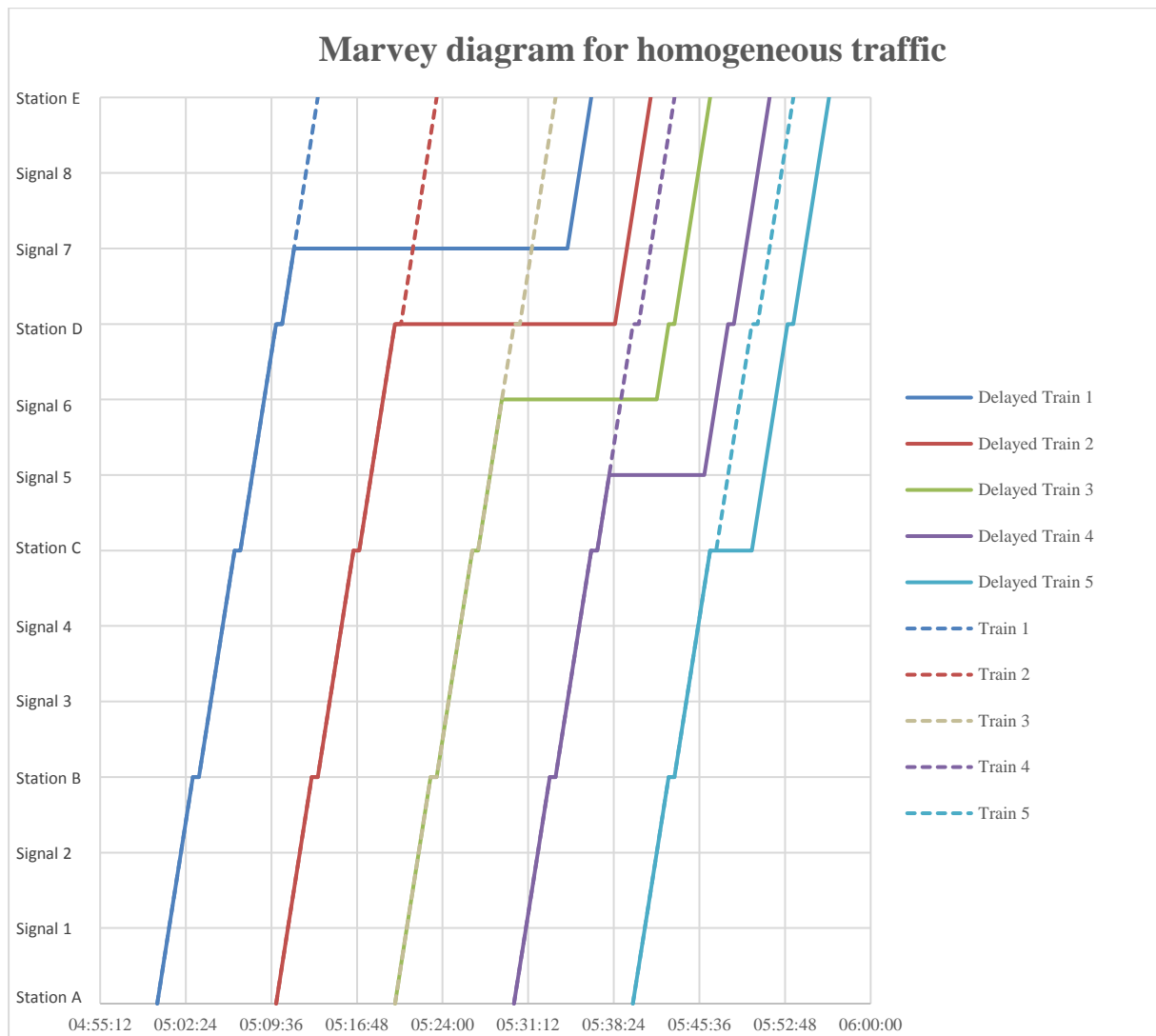


Figure 4-2: Marvey diagram for normal and delayed homogeneous rail traffic consisting of 5 trains with the primary delay occurring at Signal 7.

It can however be assumed that more delays occur during peak periods, since the train frequency is higher and thus a higher probability of failure in any part of the system and human disruption. The knock-on effect of primary delays is also greater during peak periods, since headways are shorter. In this model the average total number of delays in a day will be initially divided proportionally according to the number of trains running in peak and off-peak periods to determine the frequency of delays. During the calibration of the model, these two values will however be adjusted to compensate for limitations and assumptions made elsewhere in the model.

The last parameter that defines a primary delay is its duration. The duration of a delay depends on the type of delay. Each type has a certain probability distribution which in most cases is exponential [12]. Before a primary delay's duration can be determined, it must first be determined what type of delay is occurring. The type of delay is determined empirically by assigning a calculated percentage probability to each type and then running a random number generator at a uniform distribution for numbers between 0 and 1. The uniform distribution between 0 and 1 is then divided proportionally into intervals corresponding to the percentages calculated for each type of delay. The delay type is

thus determined as mentioned, empirically. Once the delay type is determined, the duration can be determined by the distribution assigned to that specific type.

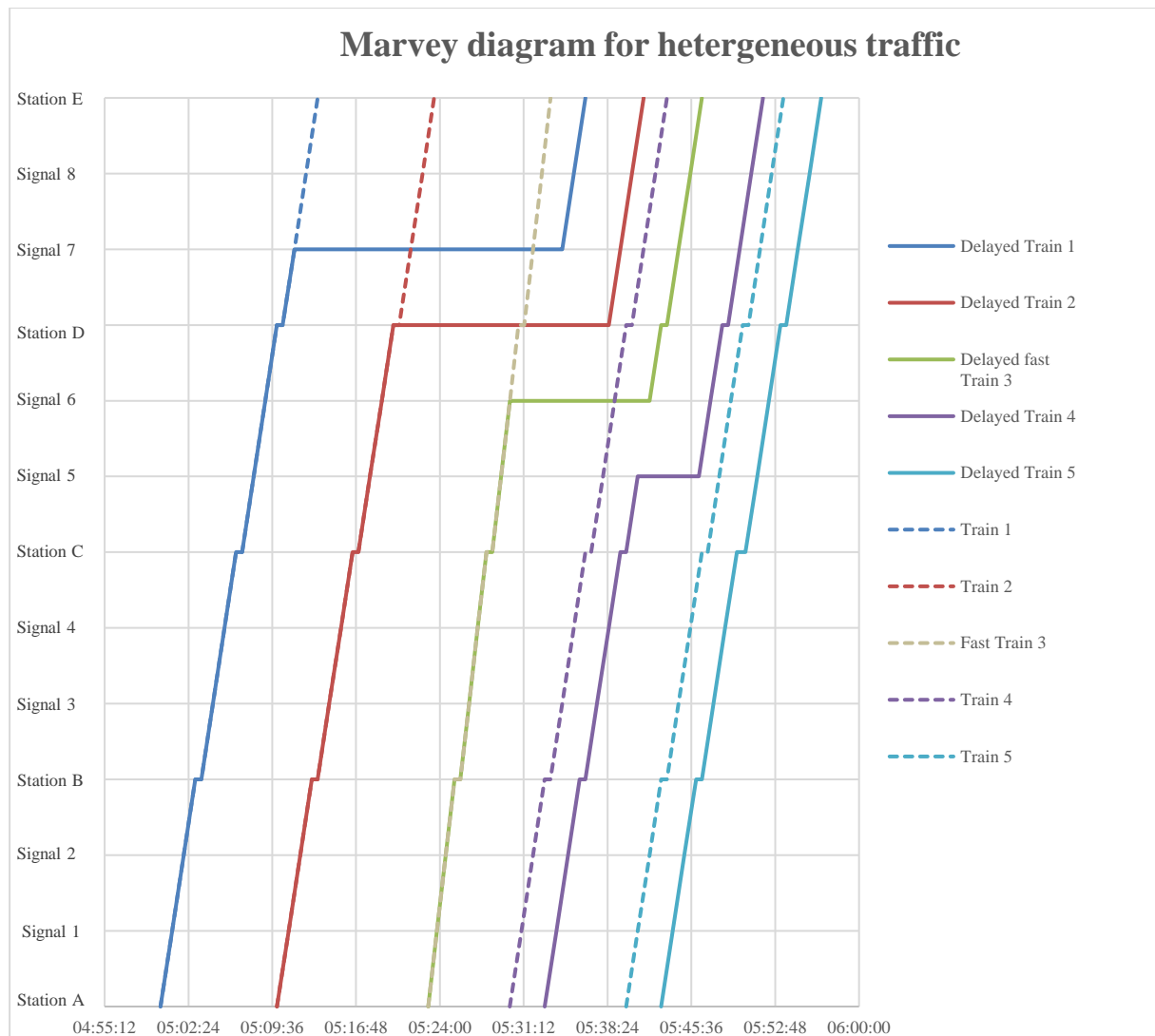


Figure 4-3: Marvey diagram for normal and delayed rail traffic consisting of 4 slow trains and 1 fast train with the primary delay occurring at Signal 7

4.1.3 Simulation software

The model can be classified as a macroscopic model since statistical data will be used to describe train behaviour and detailed train movements are simplified (reference to Section 3.3.1). Evidently, microscopic software packages such as RTC and OpenTrack were not considered. Furthermore instigating a large number of random delays over a long period of time is difficult since these packages are designed to calculate line capacity and trip times.

It was therefore necessary to use a general multi-method simulation package which will be able to capture the important aspects of train operations and the stochastic nature of an unreliable railway system. The two software packages that were considered are Simio, Anylogic. Even though there are

many other simulation software available in the market, these two packages are two of the most widely used in industry. Support and licenses were also readily available to the author.

Anylogic was chosen as the software to build this model, since it has the ability to build micro- and macroscopic rail models. Case studies of Anylogic being used to model rail operations in industry were also publically available (refer to Section 3.3.1 and 3.3.2). Even though it can be argued that Simio will also be able build the model in this study, the lack of examples and case studies of similar application in the rail environment made Anylogic the preferred choice over Simio. Table 4-1 compares the different software packages that were considered. It is clear that Anylogic is the most dynamic in its abilities and therefore further substantiates the reasoning behind choosing Anylogic as the preferred software for this study.

Table 4-1: Summary of the different simulation software packages considered for this study

Software	Modelling scope	Type of simulation	Examples in literature of application in the rail environment
OpenTrack	Microscopic	Discrete-event	Yes
RTC	Microscopic	Discrete-event	Yes
Anylogic	Micro-/Macroscopic	Discrete-event/ Agent based/ System dynamic	Yes
Simio	Micro-/Macroscopic	Discrete-event	No

4.1.4 Summary

Using Anylogic simulation software the model developed in this study will therefore produce a measure of punctuality of a rail network subjected to a variety of primary delay types. Figure 4-4 shows a basic outline of the model.

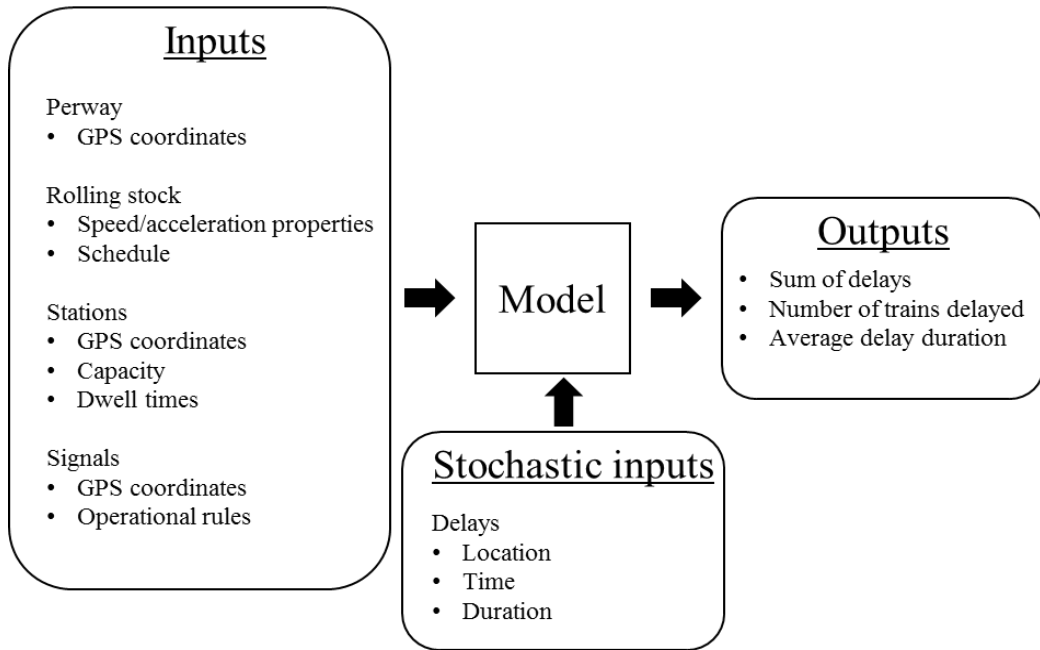


Figure 4-4: Model outline

4.2 Model – Infrastructure sub-model

Discrete event based theory is used to model the track infrastructure and the movement of trains on it. The *source block* will represent the start station from which trains will begin their journey. Thereafter a *queue block* and a *hold block* are used to represent a signal. A station is modelled by a *queue block* and a *delay block*, while the end station is the *sink block*. An *agent* or train will then start at the *source* and move through the signals and stations until it reaches the *sink*. If the signal is red, the train will be stopped at the *hold block* and wait in the *queue*. When the signal turns green again the train will proceed to the next signals until it reaches a station where it would dwell for a specified time. When the train then reaches the end station or *sink* it will exit the system. A simplified Anylogic version of this model is shown in Figure 4-5. Note that there are *moveTo blocks* as well. These represent the movement of the train in sections. From here on this sub-model will be referred to as the infrastructure model.

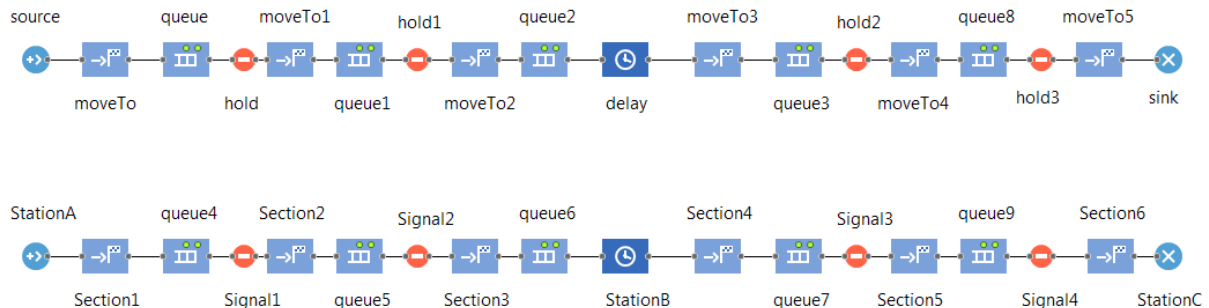


Figure 4-5: Basic Anylogic discrete event model. The top row shows the standard DE process blocks, while the bottom row shows how these were translated to rail infrastructure terms.

4.3 Model – System agent

However to capture the stochastic nature of delays and system reliability, this discrete event theory is combined with agent based theory. Two agent types are created – system agent and a train agent. The system agent determines the location, time and duration of a delay. To enforce the delay, the system agent interacts with the infrastructure model by turning a randomly selected signal red, at a calculated time for a randomly determined duration.

The system agent as seen in Figure 4-7, is either in a *Delayed State* or *Undelayed State*³. When the model starts, seed values are assigned to *t0* (time at which the system enters into its *Undelayed State*) and *delayLocation* parameters. The agent then enters into its *Undelayed State* where a random number between 0 and 1 is assigned to *delayType* to determine the first parameter of the upcoming delay when the system goes into its *Delayed State*. A Switch/Case Java function is used to facilitate an empirical distribution by means of numerical intervals proportionally sized according to the corresponding probabilities of each type of delay occurring (distributions were sourced from the observed data). For instance as can be seen in Figure 4-7, there is an 8.4% probability that a delay will be customer related. Therefore if the random number assigned to *delayType* falls within the interval between 0 and 0.084, the value of *delay* will be “Customer”. The same applies for all the other types of delays.

After the upcoming delay type is determined, another Switch/Case function is used to change any previous delayed signal to green (note that when referred to “delay signal” it means a signal that was turned red because of a delay instigated by the system agent). This code does not apply at the start of the model since no delay has been instigated yet, however after the first delay was instigated it is necessary to reset the delayed signal to green.

Next the time will be stepped forward and the time of day will be returned to determine if it falls within the peak or off-peak period. The time to the next delay will depend on which period of the day it is. The time to delay will always be shorter in peak periods since more trains run in that period and therefore increasing the probability of a delay occurring. It can be noted that *timeToDelay* is a product of delay frequency and system reliability. The *systemRel* parameter is used as a crude value to increase the whole system’s reliability by simply increasing the *timeToDelay* value by an adjustable factor. If the time to delay then expires the agent will log the time as *t1* and move into the *Delayed State*.

Another Switch/Case function is used to determine the delay duration based on the delay type determined in the previous state. It can be noted that the value for *delayDuration* is sampled from exponential distributions since it was found from data that the durations of all the types of delays are distributed exponentially. The first value in brackets is the shape factor or λ which is simply the $1/\bar{x}$. The second value is the minimum value which is 5min, since delays shorter than 5min are not accounted for.

³ For a more detailed description of the code used for the System Agent, please refer to Appendix B.

The last delay parameter that is determined is the location of the delay. One signal in every section between stations was chosen as potential delay signals, which counts to a total of 16 “delay” signals. Similar to the *delayType* parameter, the *delayLocation* value is chosen by a random number generator for values between 0 and 1. The 16 intervals are however of equal size, resulting in a uniform distribution to sample the location of the delay from. Since the location of the delay has no influence on the sum of delays as proven in Section 4.1.2, a uniform distribution is adequate. The code “`signal_8656.block()`” is a function used in *Anylogic* (the software used to build the model) to turn the specific signal to red.

After the delay has thus been instigated the code will step through time and stay in the *Delayed state* until the delay duration has expired. This will mean that any train approaching the delayed signal will stop and queue until it is turned green again. After the delay duration has expired the new *t0* will be logged and the system agent will move into the *Undelayed State* again.

The system agent model described above only accounts for one type of train agent. The train agent will be explained in Section 4.4; however it is important to understand how the system agent will account for the reliability differences in train types. In this model it will be assumed that the new trains will not experience any rolling stock related delays. Therefore if a rolling stock related delay is instigated by turning a signal red, that red signal must not apply if the oncoming train is a new train and must only apply for old trains. To account for new trains the potential delay signals in the infrastructure model was modified as in Figure 4-6. The *separator block* (“s”) is used to let all the old trains pass through “Signal5” and the new trains through “Signal5_new”. An additional *source block* was also added to function as the source of the new trains.

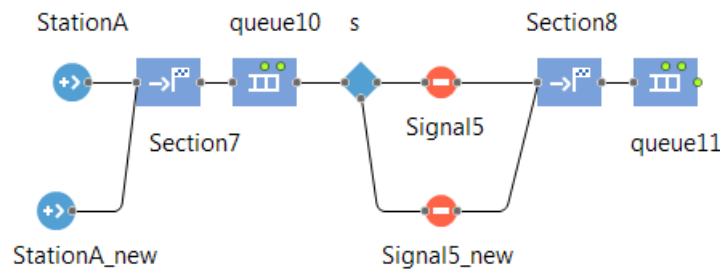


Figure 4-6: Process block arrangement to account for two train types

The algorithm shown in Figure 4-7 still applies with the inclusion of the new trains, however the code behind the process blocks related to signals in the *Delayed* and *Undelayed States* were modified. Appendix B contains the detailed code of this modification.

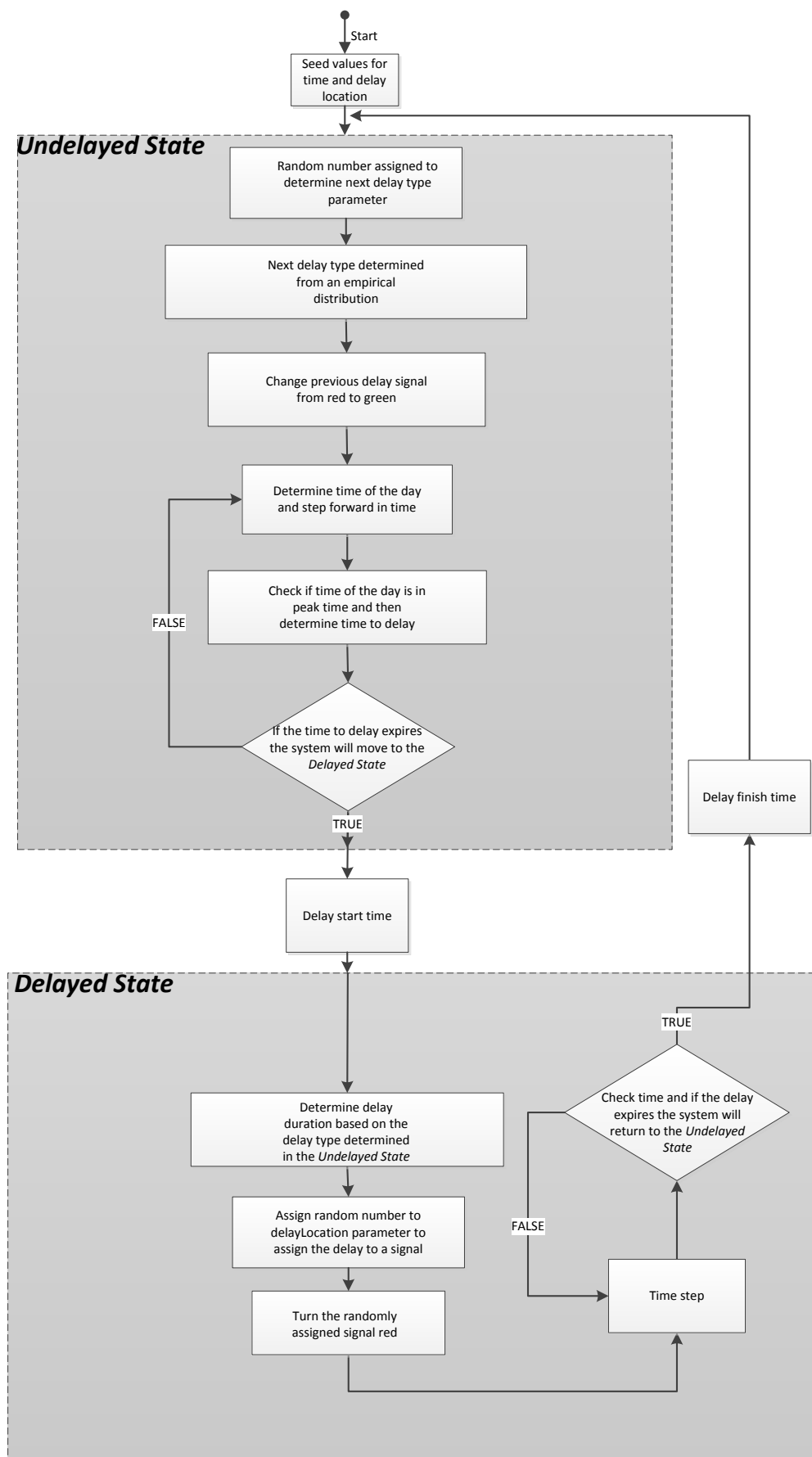


Figure 4-7: System agent algorithm accounting only for one train type

4.4 Model – Train agent

The train agent operates within the infrastructure model, but in order to incorporate acceleration, the agent has a built-in algorithm that enables the train to accelerate and cruise appropriately.

Real fixed block signalling systems commonly work on a three signal system, where a green signal will mean proceed at the specified speed limit, a yellow signal will warn the driver to drive cautiously because the next signal is red. A red signal will mean that the train must stop until the signal turns green again. When a signal is red it will mean that the section ahead is either occupied by another train or there might be maintenance occupation. This model does not take into account driver behaviour and also assumes infinite braking ability. This allows the model to use a two signal system - green and red. Furthermore the signals of this model will be used to enforce minimum headways and delays.

As mentioned in Section 4.1.2 one of the inputs to the model is speed and acceleration properties of rolling stock. To create a heterogeneous train fleet, different speed properties has to be assigned to each train type. Thus a train must be able to accelerate until it reaches a specified cruising speed and then at some point start decelerating in order to stop at either a signal or a station. However, because agents must operate within a discrete-event environment, they are only able to stop at set locations. Moreover agents will have to stop at random times if delays occur. This creates the dilemma that a train will sometimes have to stop at a signal at an unspecified and possibly unrealistic deceleration. This model then simplifies the problem by assuming infinite deceleration as previously mentioned. Further discussion on this simplification is covered later in Section 4.5.

A summary of the train agent algorithm is illustrated as a flow diagram in Figure 4-8. This algorithm is inherent to each train.

A train will depart from its starting station according schedule. If in the rare occasion the section (assuming one section min-headway between trains) ahead is occupied because of a broken train or delay-causing failure, the train will fall into a queue until the section is open. The train will then accelerate and cruise up to the next signal.

If the signal is green the train will continue cruising to the next signal or station. If the signal is red, it will be either because the section ahead is occupied, or it is experiencing a primary delay. If the section is occupied the train will again fall into a queue until the signal turns green. If the signal is enforcing a primary delay, it must first be determined what the *delayType* is. If the train type is affected by the determined *delayType* the train will be delayed for the corresponding *delayDuration*. After the delay the algorithm loops back to the signal again. If the train type is not affected by the *delayType* the train will ignore the red signal and keep on moving towards the next signal or station (refer to Section 4.3 and Figure 4-6).

If the train again arrives at a signal the same algorithm loop explained in the previous paragraph will be followed. If the train arrives at a station, it will dwell at the station for the specified duration and

then continue forward in the same way as it departed from the starting station. If however the station is the end station, the train will exit the model.

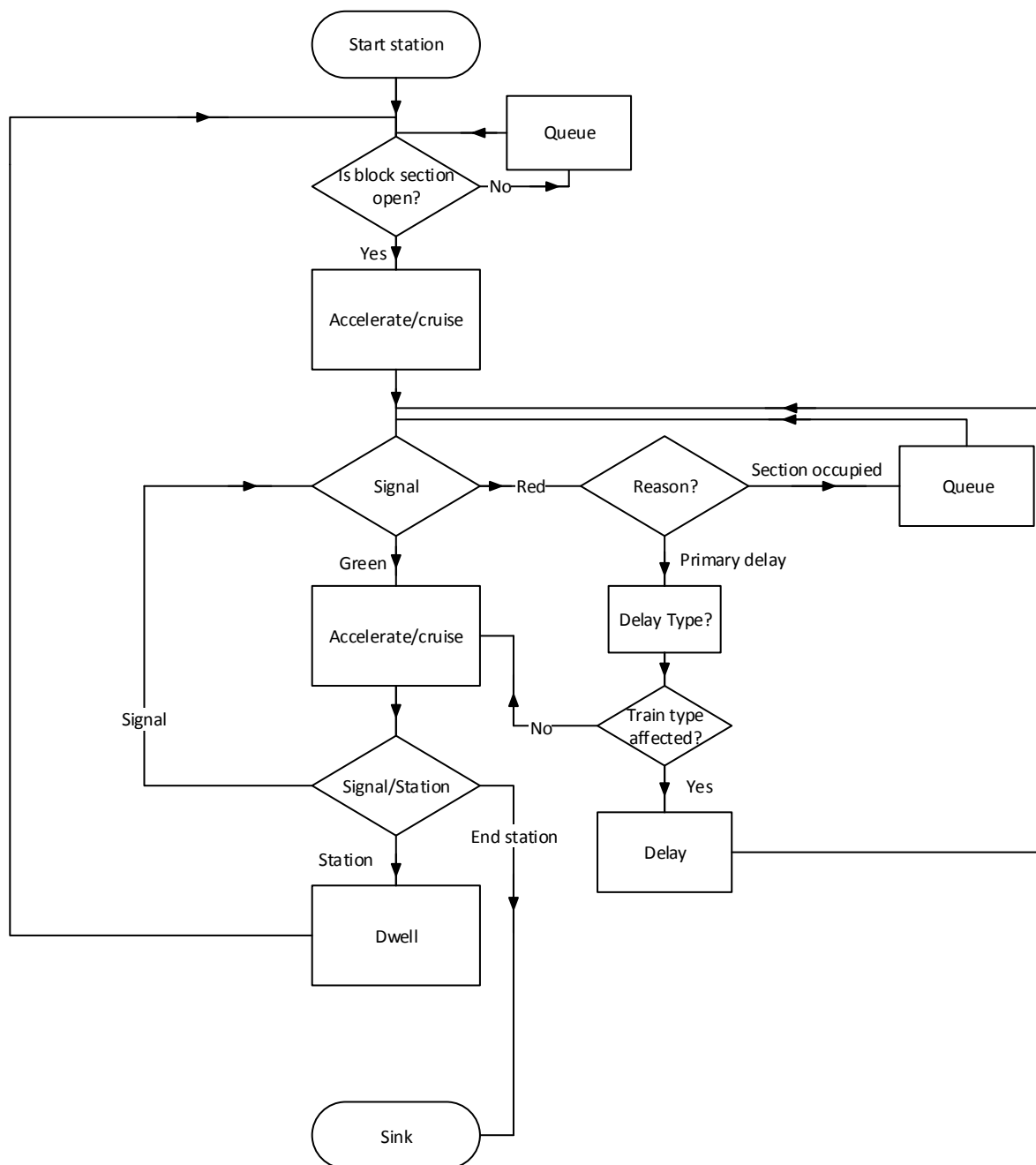


Figure 4-8: Flow diagram of the Train agent

4.5 Limitations and assumptions

This section will discuss the limitations of the model and what assumptions are made to simplify the problem into something that is computable and still representative of the reality.

4.5.1 Simulating delays

The way the model simulates delays requires calibration of the frequency of delays to ensure the correct amount of delays. As mentioned before, delays are simulated by stopping trains at signals. However the *timeToDelay* parameter is based on the model time and not on the train's itinerary. This

means a delay (i.e. a red signal) may occur, regardless if there is a train at the signal or not. It may therefore happen that a signal is blocked at time T_0 for duration of x minutes. A train may then only arrive at that signal at time T_1 and be effectively delayed for $D_p = x - (T_1 - T_0)$ minutes. It is also possible that no train arrives at the blocked signal during the *Duration* of the delay, causing a zero effective delay.

This can be accounted for by increasing the frequency of delays, which means shortening the *timeToDelay* for either or both the peak and off-peak periods. Shortening the *timeToDelay* in peak times may result in more secondary delays, because of trains running on shorter headways than in the case with off-peak periods. Since the outputs of the model are average delay duration and sum of delays, changing the delay frequency is a useful way of calibrating the model so that the model's *effective* average delays and sum of delays will match that of the real world system being modelled. The other way of calibrating would be to adjust the duration of delays. This will however be a complicated process since the delay durations are determined from probability distributions that are fitted to observed distributions from data. Adjusting these input distributions may consequently compromise the validity of the more than one input parameter. Adjusting only the frequency of delays however only compromises the validity of one input parameter.

4.5.2 Acceleration and deceleration properties

As mentioned in Section 4.4, delays are enforced by blocking a random signal for a random duration. When this delay occurs it is not based on the train's schedule, but rather on a set frequency depending on the time of day. Thus as long as a train is approaching a signal, the train has to stop at the signal if it turns red, regardless of the distance between them. This means that there can be no limit to the train's deceleration capability. In reality if a signal turns red and the train is too close to stop before the signal, the train is allowed to move past the signal while breaking. As mentioned in Section 4.4, real railway signals have a yellow signal as well to warn the train that the next signal is red. This is however not possible to model with discrete event modelling, since the entities are only able to stop at specific points in the model space and not able to move past a point because of momentum. It will thus be assumed that the train decelerates at an infinite rate (i.e. the train will stop immediately regardless of its current speed). The effect of this assumption is that the model train will cover its scheduled distance in a shorter time than the real train with the same acceleration and cruising speed properties. To compensate for this assumption, the acceleration and cruising speeds of the model train are reduced so that the time won by instant deceleration is lost in slower acceleration and cruising speeds. Figure 4-9 shows a typical speed profile for a train – accelerating from Station A until a cruising speed is reached. The train will cruise until it has to decelerate again to stop at Station B. The model train as mentioned will in this example have to accelerate slower to end up at Station B at the same time as the real train since it has no deceleration curve. Drivers may also simply remove the power supply to the traction motors and allow the train to coast and gradually decelerate until it is necessary to apply

brakes. The coasting regime however is very dependent on the driver's judgement and preference, and thus difficult to model. Coasting will thus not be considered in the model.

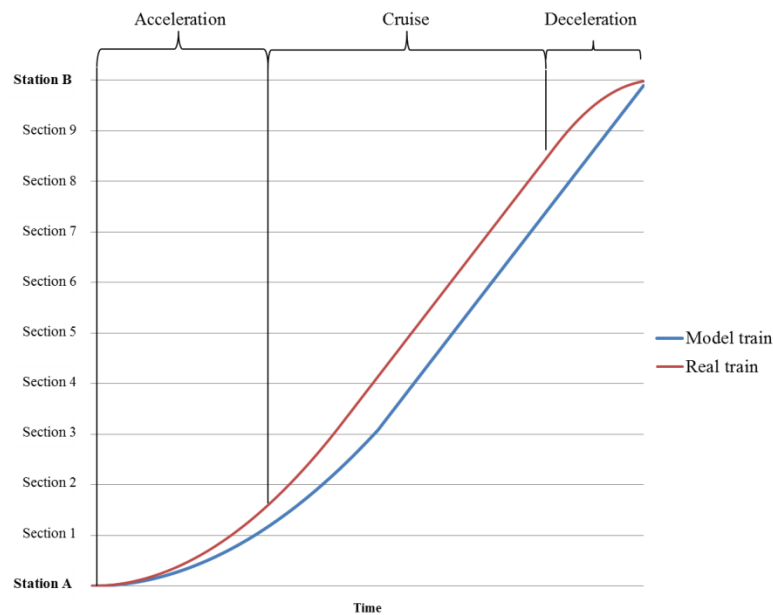


Figure 4-9: Model acceleration curve vs real acceleration curve

4.5.3 Train passing

In the rail environment train passing usually refers to trains running in opposite directions on a single line, and having to pass each other by use of passing loops. The line considered in this model however is a double line and thus train passing refers to trains running in the same direction having to pass each other.

Therefore, considering one line on which two trains with different speed characteristics are running in the same direction, two scenarios are possible. The first is when the slower train follows the faster train – operationally this is not a problem because the distance between them will grow with time. However the second scenario is when the faster train follows the slower train – this is an operational problem since the faster train will get caught behind the slower train. This can be prevented by proper scheduling, but schedules don't take into account random delays. Passing in the same direction will allow the faster train to get ahead of the slower train. Passing on track sections however, requires special operational effort for real train networks, and is usually applicable to large disruptions. This model focusses on small and medium length disruptions, thus track section passing falls outside the scope of this model.

In this model trains will not be able to pass each other in any way. It will be assumed that all train types stop at all stations and have the same dwell-times. Passing therefore in stations is also not possible. This assumption limits the possibilities when creating scenarios for the model.

4.5.4 Peak and off-peak delay events

As mentioned in Section 4.3 the frequency of delays are dependent on the number of trains which in turn is determined by the time of day. Now since it is fixed frequencies for the two different periods, one would think that there are a fixed amount of delays each day. However with close investigation of the algorithm illustrated in Figure 4-7, it can be noticed that this is not necessarily true. Primary delays are calculated one at a time, and also occur one at a time. This means the primary delays cannot overlap or run concurrently⁴, because the system is either in its *Delayed State* or *Undelayed State*. The next *timeToDelay* is only determined when the system enters the *unblock* state. When *timeToDelay* is determined, it is not based on the time of day; instead it is only the value of *timeToDelay* that is based on the time of day.

In reality primary delays can occur concurrently and a rail network system is much more complex having three signalling states (i.e. green, yellow and red). To make this model computable, the system was unfortunately simplified to just two states (green and red). The calibration and validation of the model is therefore a crucial step. To get the model to give relatively realistic outputs it will be necessary to adjust some of the input parameters to make up for the simplifications and assumptions. The parameters to be configured will depend on the nature of the system being modelled and should be done with care and consideration of the integrity of the model.

4.6 Summary

This chapter explained the structure and development of the model. It stated the model outputs, inputs and overall structure. The structure consists of two agent types namely: System agent and Train agent. The algorithms behind these agent types were explained and also the accompanying limitations and assumptions that were made in order for the model to work.

The application of this model on the case study described in Chapter 2 will be covered in Chapter 5.

⁴ In the reality primary delays may occur concurrently. However because the model controls delays centrally instead of locally, the model is limited to only one delay at a time.

5. Case study model

Before the new trains can be modelled into the network the system must first be modelled in its current state, and thereby compare the model outputs to that of real data. A conclusion as to if the model is valid can then be drawn. In this study the line between Chris Hani and Cape Town will be modelled which will only include traffic in one direction (i.e. direction up). The 39 km line consists of 18 stations, 57 signal blocks and carries 291 trains per week in one way. Figure 5-1 illustrates the network. It is important to note that trains can run from Chris Hani and Kapteinsklop to Cape Town via Pinelands or via Mutual. This model however only includes the trains running via Mutual. The trains running via Pinelands do not run on the same tracks as the Mutual trains, and thus will not have an effect on the Mutual traffic. Some trains from Kapteinsklop however also run via Mutual to Cape Town and are thus included in the model. The Sarepta and Bellville trains also run on separate lines to Cape Town, and will also be excluded from the model.

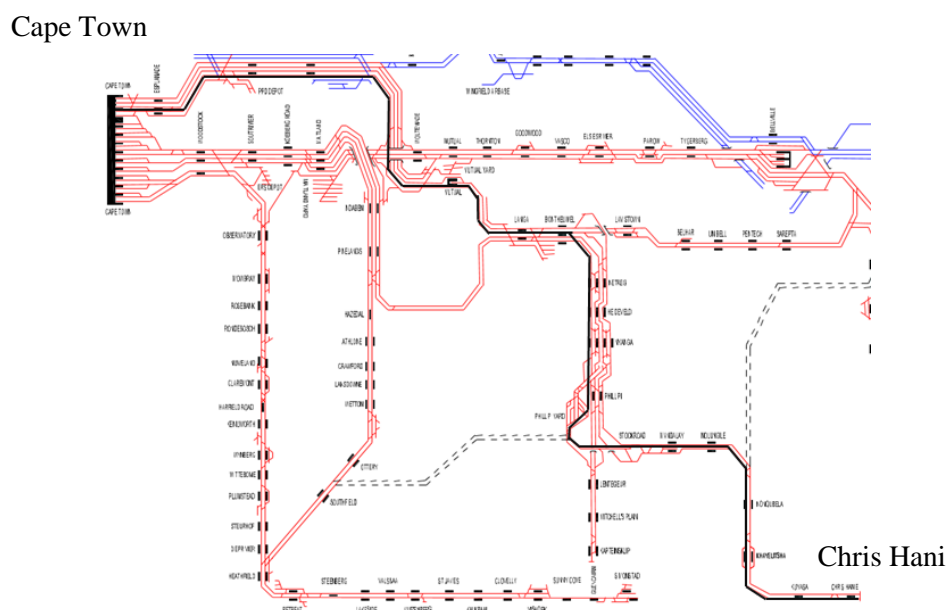


Figure 5-1: Chris Hani to Cape Town network diagram

This chapter will show how the model developed in Chapter 4 is applied to this case study. In Section 5.1 and 5.2 the input data collection and processing will be discussed. Sections 5.3 and 5.4 will cover the validation process and how the relevant software was used to simulate the model.

5.1 Inputs

Anylogic simulation software was used to model the network. The input parameters and variables listed in Figure 4-4 were processed to be compatible with the software. This Section will discuss that process.

5.1.1 Perway

In order to model and represent the route from Chris Hani to Cape Town station, a GIS map was used. The route was then laid out by placing GPS points along the track as visually observed. These points were then connected to form the route on which the trains will run.

5.1.2 Rolling stock

Since the reliability of the train types are modelled inputs under delays, the only inputs concerning rolling stock are speed and acceleration properties and the schedule. In the “as-is” model only one train type will be considered, namely *old train*. *Old train* will represent the 5M2A train set which is the most common type operating on the Western Cape network. The 10M train sets which also operate on the Western Cape network, uses the same drive system as the 5M2A and thus can be assumed the same speed properties. The 8M trains use a different drive system, however only one currently operates on the Western Cape network according to Mr. Robert Venter from *PRASA*. It will therefore be assumed that *old train* will accelerate at a maximum of 0.35m/s^2 and cruise at a maximum of 80km/hr. These two parameters can easily be adjusted to calibrate the train’s traveling time between the Chris Hani and Cape Town. This calibration process will be further discussed in Section 5.3.

Anylogic 7.1 has a *Rail* library with which train movement can easily be modelled. However with this version it is not possible to use the *rail* library to model on the GIS map. It was therefore necessary to use the *Process Modelling* library which is able to model on the GIS map. The *Process Modelling* Library can however only move the agents at a constant speed and acceleration and deceleration are instant. Therefore an acceleration algorithm had to be developed and embedded into the *old train* agent. The algorithm is as follow:

The agent departs from its stationary point with a speed value of 1. This is just a seeding value to kick-start the algorithm. The model works in increments of 1 second, and thus to accelerate the agent, its speed is multiplied by (1+acceleration value) every second. The speed will increase until the cruising speed is reached, after which the agent’s speed will stay constant until it has reached the next discrete event. The discrete event will serve as the next stationary point at which the algorithm will start again. Figure 5-2 illustrates the flow diagram of the acceleration algorithm. Technically every second in the model is a discrete event; however a “discrete event” in this algorithm refers to a train having to stop at either a signal or station. As mentioned in Chapter 4 the trains in this model stop instantaneously and thus have an infinite deceleration rate. The algorithm is coded in *Java script* and implemented by *Anylogic* to determine the movement of the *Type train* agents.

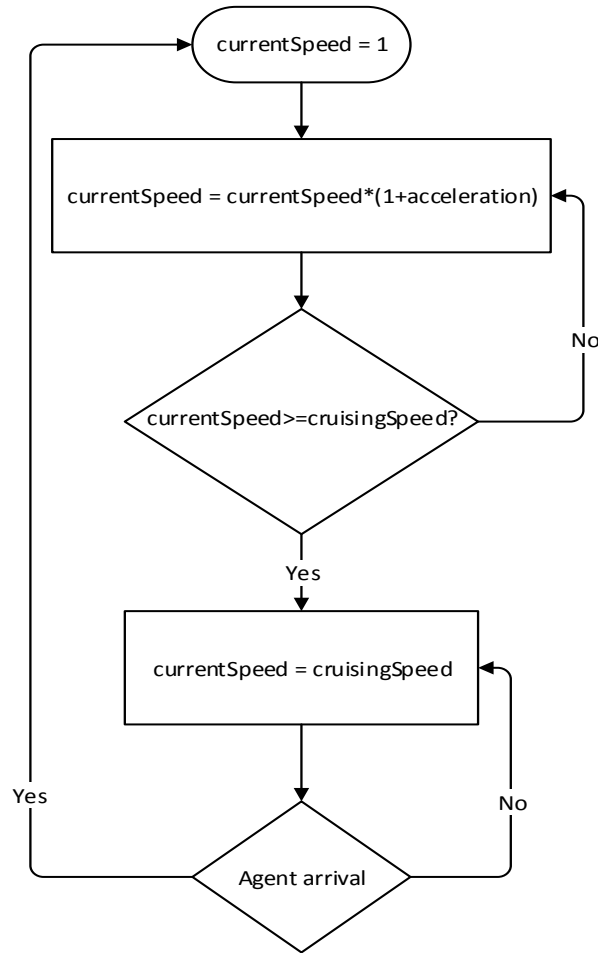


Figure 5-2: Train agent's speed and acceleration algorithm

The second input under rolling stock is the departure schedule from the starting stations. Trains will depart from Chris Hani, Khayelitsha and Phillipi stations. Most of the trains in the model starting from Phillipi are trains that in reality start their route from Kapteinsklop station. Because the track between Phillipi and Kapteinsklop only has two stations, it is left out of the model. The influence this piece of tracks has on the rest of the network is negligible, however as mentioned the trains running from Kapteinsklop will be included in the schedule for the trains starting from Phillipi. All the trains that share the same track are included in the model. The schedule was received from *PRASA's* train operations office. A copy of the schedule can be seen in Appendix A1.

It is important to note that the schedules are only departure schedules for trains at their starting stations. The arrival schedule at their end stations will be used to compare the model's trip time to that of the real scheduled trip time. This comparison will then be used to calibrate the model. Section 5.3 will elaborate on the calibration process.

5.1.3 Stations

The details concerning the stations are important in describing the model. The locations are also determined by GPS coordinates retrieved from *Google Earth*. Trains will stop at every station in the model except for Paardeneiland station, which only serves as a depot.

Stations in the network have different platform capacities; however in this model it will be assumed that all stations have only one platform and one line in each direction. This means that there is no way trains in the model can overtake one another. As mentioned before in reality if a train is delayed at a station, following trains could pass if there is more than one platform. In the model however the locations where the primary delays will occur are chosen not to be at stations and therefore eliminate the need for trains to pass at stations. This arrangement will in fact cause more secondary delays, and thus the necessity to again calibrate the model with delays. Section 5.4 will cover the calibration process that includes delays in more detail.

The third parameter concerning stations is the dwell times of trains. In most cases the dwell time is 30 seconds, except for Mandalay, Philippi and Bonteheuwel station where it is 60 seconds. Paardeneiland as mentioned has no dwell time. In reality trains don't always dwell according to the scheduled time because of several reasons pertaining to delays which will be discussed later. A train driver thus dwelling at a station longer than scheduled will log the time lost as a delay. The model takes this into account by including it as a delay type. However it may happen that a train driver will want to catch up time lost by shortening his dwell time. This is not specifically taken into account by the model. The effect of this limitation on the output is negligible since a train driver only logs a delay at his end station. If he thus made up lost time by dwelling shorter at stations it would only reduce the total delay logged at the end station. Trains rarely arrive early at stations. A summary of the station inputs are given in Table 5-1.

Table 5-1: Station inputs

Station	Location		Capacity	Dwell time (s)
	Latitude	Longitude		
Chris Hani	34° 3'17.34"S	18°42'38.75"E	1	Start station
Kuyasa	34° 3'17.48"S	18°41'36.13"E	1	30
Khayelitsha	34° 2'52.50"S	18°40'14.68"E	1	30
Nonkqubela	34° 1'36.83"S	18°39'47.14"E	1	30
Nolungile	34° 1'1.02"S	18°38'56.34"E	1	30
Mandalay	34° 1'8.20"S	18°37'28.53"E	1	60
Stock Road	34° 0'51.23"S	18°36'22.62"E	1	30
Philippi	34° 0'48.33"S	18°35'4.33"E	1	60
Nyanga	33°59'33.95"S	18°33'36.11"E	1	30
Heideveld	33°58'11.35"S	18°33'42.70"E	1	30
Netreg	33°57'9.45"S	18°33'49.30"E	1	30
Bonteheuwel	33°56'31.03"S	18°33'0.07"E	1	60
Langa	33°56'20.40"S	18°31'47.77"E	1	30
Mutual	33°55'18.74"S	18°30'47.71"E	1	30
Ysterplaat	33°55'11.80"S	18°28'37.08"E	1	30
Paardeneiland	33°55'21.12"S	18°27'59.62"E	1	0
Esplanade	33°55'25.81"S	18°26'43.52"E	1	30
Cape Town	33°55'25.14"S	18°25'38.61"E	1	End station

5.1.4 Signals

To determine the signal locations *Softtime* data was retrieved from *PRASA*'s infrastructure department at Salt River depot which indicated the distance from each signal relative to Cape Town station. Because the *Google Earth* images did not clearly show each signal, the distances between signals was calculated and then by use of *Google Earth* the locations of each signal were estimated. The coordinates of the signals were then entered into *Anylogic* to pin point where each signal is located.

Since the modelled network does not include any junctions or switches, the operational rules applied are rather simple;

- On track sections there must always be one block section open between trains.
- At stations there must be a block section clear on either side of the station. In effect there must be two block sections open between trains before and after a station.

Figure 5-3 shows the GIS map with all the signal and station locations. This map was used to animate the simulation.

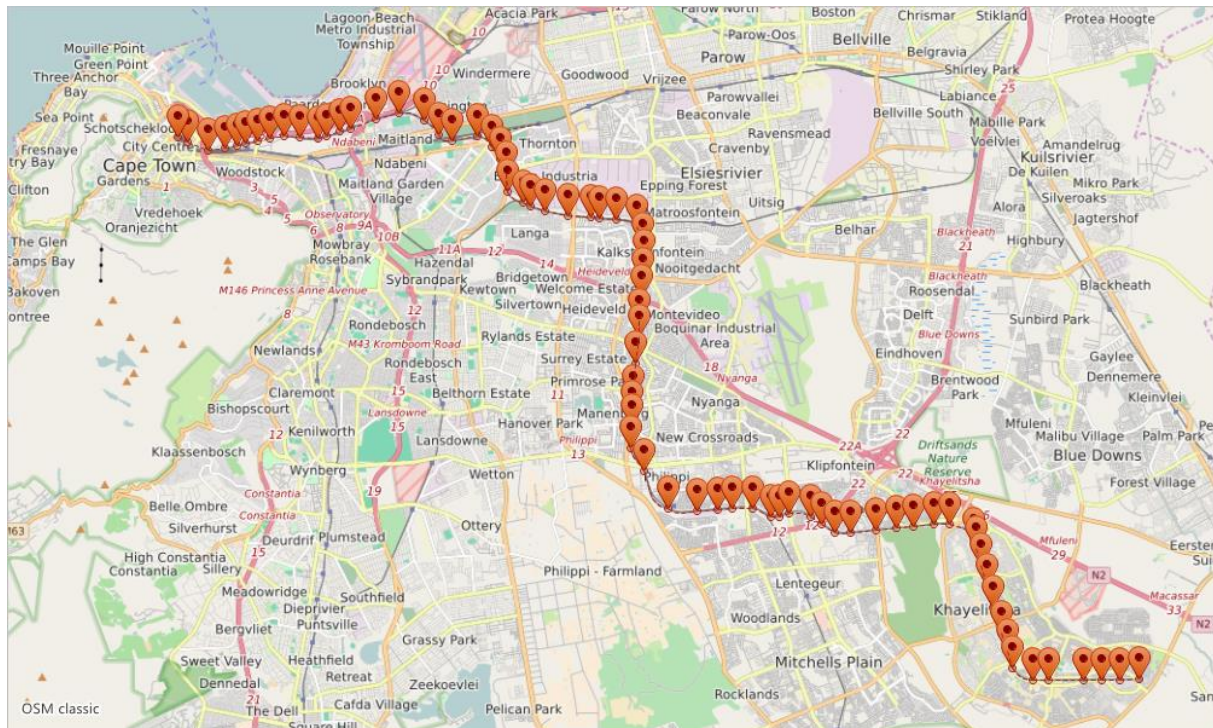


Figure 5-3: GIS map of the station and signal locations between Chris Hani and Cape Town stations

5.2 Stochastic inputs

The stochastic inputs are added dynamically throughout the run of the model. This section will cover the process of how the randomness of the inputs was determined. As shown in Figure 4-4 these inputs include the location, time and duration of the primary delays. The data mined and processed for this model is delay data received from *PRASA*'s operations office for the months April 2015 to September 2015. Figure 5-4 shows an example of an extract from the raw delay data received from *PRASA*.


Filtered Delays Reported: 2015-07-01 to 2015-07-15									
 2015/10/02									
Delays: 598.0 Canc: 41 Minutes: 12747 Short KM: 0									
Cape Metrorail									
Departments									
Delays: 407.1 Cancel: 41 Minutes: 10505 Short KM: 0									
Customer Services									
Delays: 35.1 Cancel: 0 Minutes: 547 Short KM: 0									
Passenger Related - Overcrowding of trains									
Delays: 35.1 Cancel: 0 Minutes: 547 Short KM: 0									
DelayID	Date	Tim	TrnNo	Code	Perc	Minutes	KM	Place	EventID
367236	2015-07-01	6:00	9918	DEL	50%	6	0	Cape Town Station	73975
Description TRAINS DELAYED DUE TO LACK OF CAPACITY									
367245	2015-07-01	6:00	9404	DEL	46%	6	0	Cape Town Station	73975
Description TRAINS DELAYED DUE TO LACK OF CAPACITY									
367469	2015-07-01	6:00	9951	DEL	27%	11	0	Khayelitsha Station	73975

Figure 5-4: Extract from the delay data received from PRASA in Excel format

5.2.1 Location

As seen in Figure 5-4 a “Place” is indicated as to where the delay occurred. However according to Mr. Jacques Carstens at PRASA’s operations office, those locations do not necessarily indicate where exactly the delay occurred. Drivers many times simply log the delay under one of the major stations along the route. The only advantage of having accurate location details would be that a more representative animation of the model could have been made. Otherwise the location of the delay is not important as explained in the Chapter 4.

5.2.2 Time

In Figure 5-4 it can be noted that a time of delay is indicated. Figure 5 5 shows the distribution of these times of delays sampled over a 6 month period compared to the number of trains scheduled for each hour in the day. The graph makes sense if one looks at the peak period of the day (05:00-08:00). Note that this is the profile for the trains moving up from the sub-urban areas to the CBD in Cape Town. For the down direction the peak period will be between 17:00-19:00 when passengers return home. During peak periods train frequency is the highest, and thus the probability of a failure causing a delay to occur is also the highest. However 42% of the data entries are logged for the hours between 00:00 and 01:00. This is not an operational period of the day and cannot be regarded as reliable data entries. According to Jacques Carstens from PRASA, signalling boxes are many times vandalised and thus components necessary to capture the correct time of a delay are damaged in such a way that a default time of 01:00 is captured in many instances.

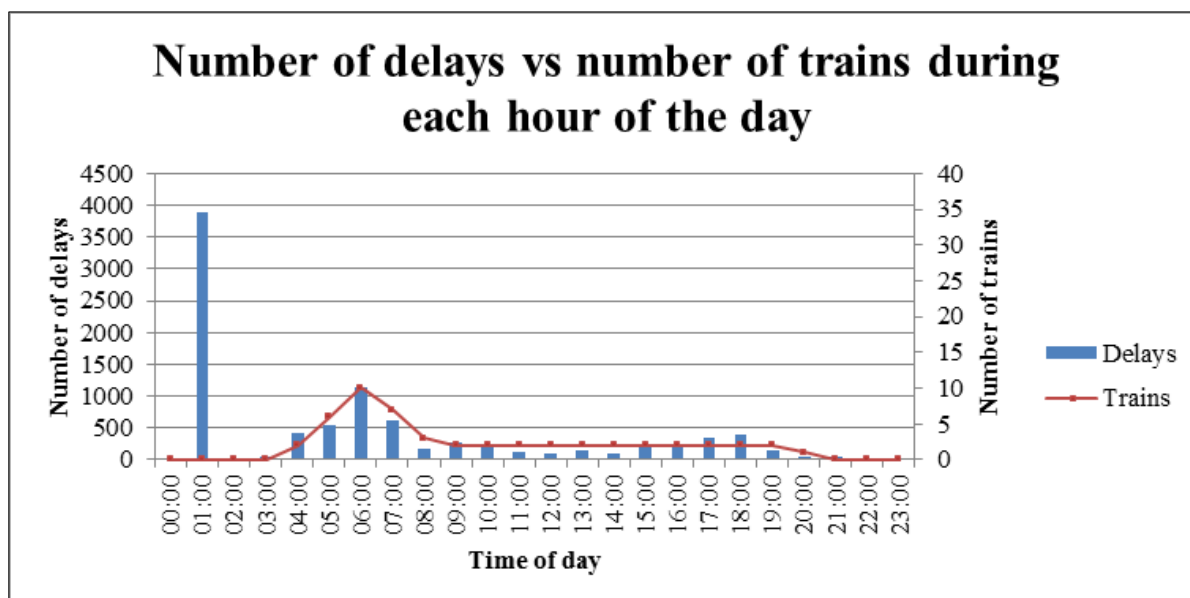


Figure 5-5: Distribution of the number of delays and number of trains during each hour of the day, sampled over 6 months only for trains running up (i.e. Chris Hani to Cape Town)

Nevertheless, since the integrity of the data can be questioned, the frequency of primary delays will be adjusted proportionally according to the number of trains running. Forty six percent of the trains run between the hours 5:00 and 8:00 and on average 7.78 primary delays occur per day calculated over a 6 month period (1 April 2015 – 30 September 2015). Assigning 46% of the delays per day to the 3 peak hours results in a delay frequency of 50.27min. The remaining 54% of the delays are assigned to the off-peak hours giving a delay frequency of 185.71min.

As mentioned before, only one primary delay is simulated at a time, and therefore the system is either in the *Delayed* state or *Undelayed* state. For this reason the system will not necessarily produce a delay every 50.27 or 185.71 minutes, but these frequencies will only be effective when the system is in the *Undelayed* state (reference to Figure 4-7).

5.2.3 Duration

As mentioned in Chapter 4, the duration of a delay will depend on the type of delay. The types of delays are grouped under the different departments of PRASA. When a train driver logs a delay, he must indicate what the cause of delay was. A meeting is then held every day by the managers of all departments to decide which department is to take the blame for the delay. The table in Appendix A2 shows all the different causes of delays. These delays are grouped under the relevant departments of PRASA. Table 5-2 shows a summary of these primary delays that occurred in the recorded 6 month period. Note that since there are such a vast number of different delay types, for simplification delay types will be named under the responsible department for the rest of this study. Also note that “*Speed restr.*” refers to temporary speed restrictions. Even though it is not necessarily a “random” event, it is logged as a delay event, and the schedule was not changed to accommodate the speed restrictions. Furthermore not all trains were delayed meaning that the speed restrictions were not applied for the whole 6 months. These temporary speed restrictions are usually enforced on sections where the track

condition is such that a train may derail when traveling at normal speeds. Because track failures and maintenance thereof are not necessarily predictable events, temporary speed restrictions will be modelled as random events.

The primary delay was assumed to be the delay logged first (refer to Figure 5-4) under an *EventID* number. The other delays logged under the same *EventID* were assumed to be the consequential delays. Personnel at PRASA could not confirm that these assumptions were valid, and neither could they give the correct interpretation of the data set.

Table 5-2: Summary of the primary delays under each department for the 6 month period

Primary delays	Overall	Customer services	Rolling stock	Protection services	Perway	Speed restr.	Signals	Other
Average duration [min]	14.4	8.0	17.5	17.4	12.8	8.7	13.5	17.4
Standard deviation [min]	14.0	3.1	15.2	13.9	7.5	2.9	12.3	22.7
Count [.]	1424	120	411	172	143	204	192	182
Trimmean (95%) – λ [min]	10.0	7.0	11.7	12.5	10.0	8.0	10.0	12.0
Average minutes to delay	131	1555	454	1085	1305	915	972	1025
Probability	100%	8%	29%	12%	10%	14%	13%	13%
Shape factor ($1/\lambda$)	0.10	0.14	0.09	0.08	0.10	0.13	0.10	0.08
Total minutes	20448	962	7174	2996	1825	1770	2589	3162

The type of delay is determined by the model by use of the empirical probabilities shown in Table 5-2. These probabilities were simply calculated by dividing the number of delays of each type by the total number of delays. A random number generator was run for numbers between 0 and 1. This range was divided into intervals proportional to the probabilities indicated in Table 5-2. Each time the generator produces a number, the type of delay will be determined by the interval in which the number falls. Note that these are values pertaining to only the primary delays that will be used as input to the model. The model output will include all the delays (i.e. primary and secondary) and therefore the number of delays will be much more.

The duration of all the types of delay showed exponential distributions. Therefore a shape factor, λ , was calculated for each type to describe the estimated exponential curve (generally expressed as in Equation 3) using the 5% trimmed mean. Because the data showed a large number of outliers the average delays had to be trimmed by 5%⁵.

$$y = \lambda e^{\lambda(-x)} \quad (3)$$

⁵ It was found extremely difficult to get useful outputs from the model when these outliers were included as inputs since they caused the model to be “too random” and to produce outputs that could not be validated in the form of a goodness-of-fit test. Forty four delays out of the 1424 primary delays, ranging between 50 and 125 minutes, were therefore trimmed to give a 5% trimming on the average delay duration. This trimming can further be justified by the fact that the model does not wish to simulate the exceptional delay events but rather the common events. Simulating exceptional events would require a separate study and possibly different assumptions and modelling method.

As an example of how these estimated delay duration distributions were validated with the observed data, refer to Figure 5-6 which shows the observed and estimated cumulative distributions for Rolling stock failures that caused a primary delay⁶. According to Law & Kelton [44] the Kolmogorov-Smirnoff (K-S) goodness-of-fit test is most appropriate when comparing cumulative distributions, and therefore the (K-S) test was used to validate the estimated distributions. The procedure for Rolling stock delays went as follows:

1 Null Hypothesis – H_0

The observed data have a theoretical cumulative distribution $F(X)$ with mean $\mu = 17$ and $\sigma^2 = 230$.

2 Calculation of right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n

$$D_n^+ = \max_{0 \leq i \leq n} \{F_n(X_i) - F(X_i)\} \quad (4)$$

$$D_n^- = \max_{0 \leq i \leq n} \{F(X_i) - F_n(X_{i-1})\} \quad (5)$$

$$D_n = \max\{D_n^+, D_n^-\} \quad (6)$$

Table 5-3: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for delay durations of rolling stock related delays

Number of delays per week						
		Observed		Model Estimate		K-S statistic
	X_i	Frequency	$F(x)$	Frequency	$F_n(x)$	D_n
X_1	0	0	0.003	0	0	0.003
X_2	5	1	0.197	139	0.174	0.023
X_3	7	77	0.348	118	0.322	0.026
X_4	9	60	0.495	100	0.447	0.047
X_5	11	58	0.571	85	0.554	0.017
X_6	13	30	0.641	72	0.644	-0.003
X_7	15	28	0.705	61	0.721	-0.017
X_8	17	25	0.747	52	0.786	-0.039
X_9	19	17	0.813	44	0.841	-0.028
X_{10}	21	26	0.848	37	0.888	-0.040
X_{11}	23	14	0.874	32	0.928	-0.054
X_{12}	25	10	0.907	21	0.955	-0.048
X_{13}	30	13	0.934	14	0.972	-0.038
X_{14}	35	11	0.952	9	0.984	-0.032
X_{15}	40	7	0.980	6	0.991	-0.012
X_{16}	45	11	0.992	4	0.997	-0.004
X_{17}	50	5	1.000	3	1.000	0

⁶ Delays less than 5 minutes are usually not logged, however there are instances in data found where the driver may log a 3 or 4 minute delay. Regardless, the model is programmed to only instigate delays larger than or equal to 5 minutes.

3 K-S goodness-of-fit test

As seen in Table 5-3 the supreme value (coloured in red) for D_n is 0.054. According to Law & Kelton[44] to compare D_n to the critical value - $c_{1-\alpha}$ - the following test condition applies:

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n > c_{1-\alpha} \quad (7)$$

For $n = 396$, $1 - \alpha = 0.975$ and $c = \mathbf{1.48}$ the result is:

$$1.08 < \mathbf{1.48}$$

Therefore there is no sufficient evidence to reject the H_0 , and estimated exponential distribution can be regarded as a good fit to the observed data. This validation process was followed with all the input distributions.

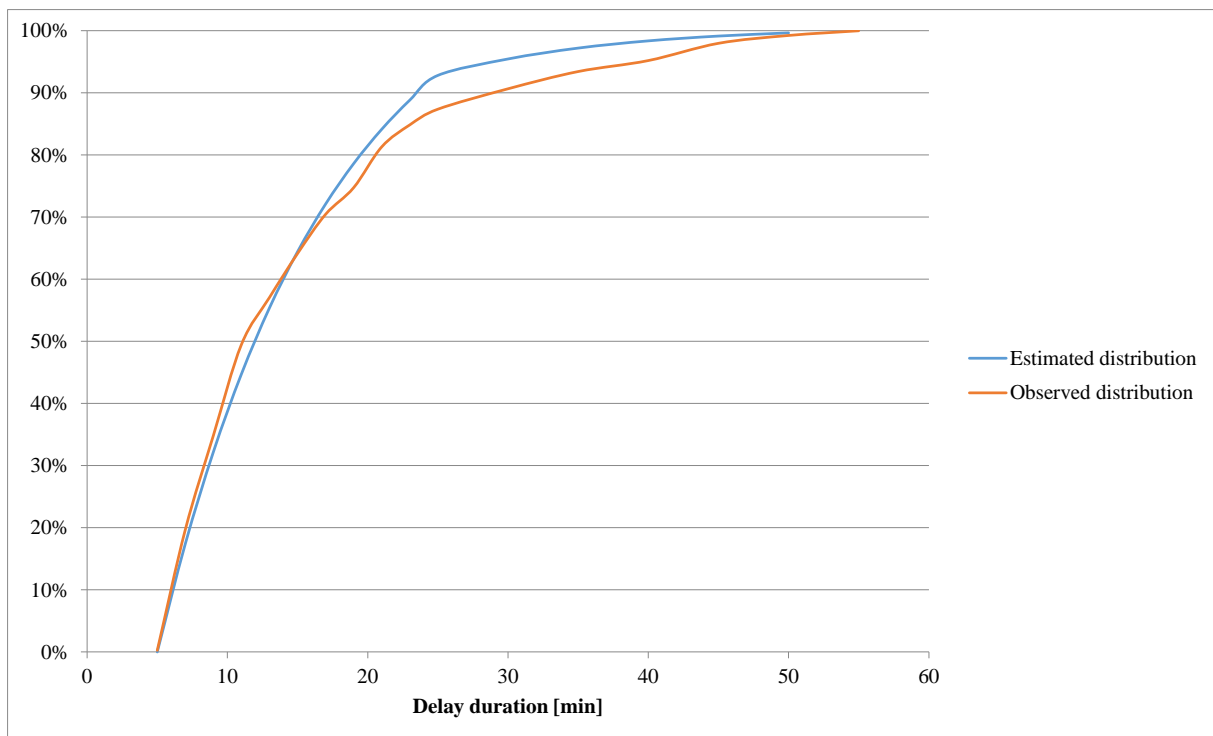


Figure 5-6: The observed and estimated cumulative distributions for rolling stock related primary delays.

5.3 Validation without delays

The model validation phase has two parts. The first part is to calibrate the train's average speed and acceleration so that the trains in the model will have the same trip time as the scheduled trip time. Of course this means that the trains cannot be delayed, and they should run as in an ideal world.

The table in Appendix A1 shows the schedule and trip times for all the trains running on a weekday. It must be noted that there is a difference in trip times for trains running the same route. For instance Train 1 departing from Chris Hani travels 1:03:00 to Cape Town, but Train 4 travels the same distance in 01:01:00. The same can be said of the Khayelitsha and Kapteinsklop routes. To simplify this complication the model trains are calibrated to travel at the longest trip times scheduled. This means

that the Chris Hani trains will travel 1:03:00, Khayelitsha trains 00:53:30, and Kapteinsklop trains 00:59:00. The difference between the scheduled and modelled trip times are illustrated in Figure 5-7. Note that the values are rounded to the nearest minute and are all on the negative side of the scale.

The maximum acceleration and speed of the 5M2A, is 0.4m/s^2 and 55km/hr . respectively, according to Mr. Robert Venter from PRASA. The model calibration thus used those values as a start and gradually reduced them until trip times corresponded to those mentioned earlier. The final calibrated acceleration and speed values are 0.21m/s^2 and 50km/hr . respectively. The much lower acceleration value can be explained by the fact that infinite deceleration was assumed and that these values are average speed values. Generally train speeds are restricted by track geometry (such as curve radii and elevations) in some sections, thus they cannot not always run at maximum speed.

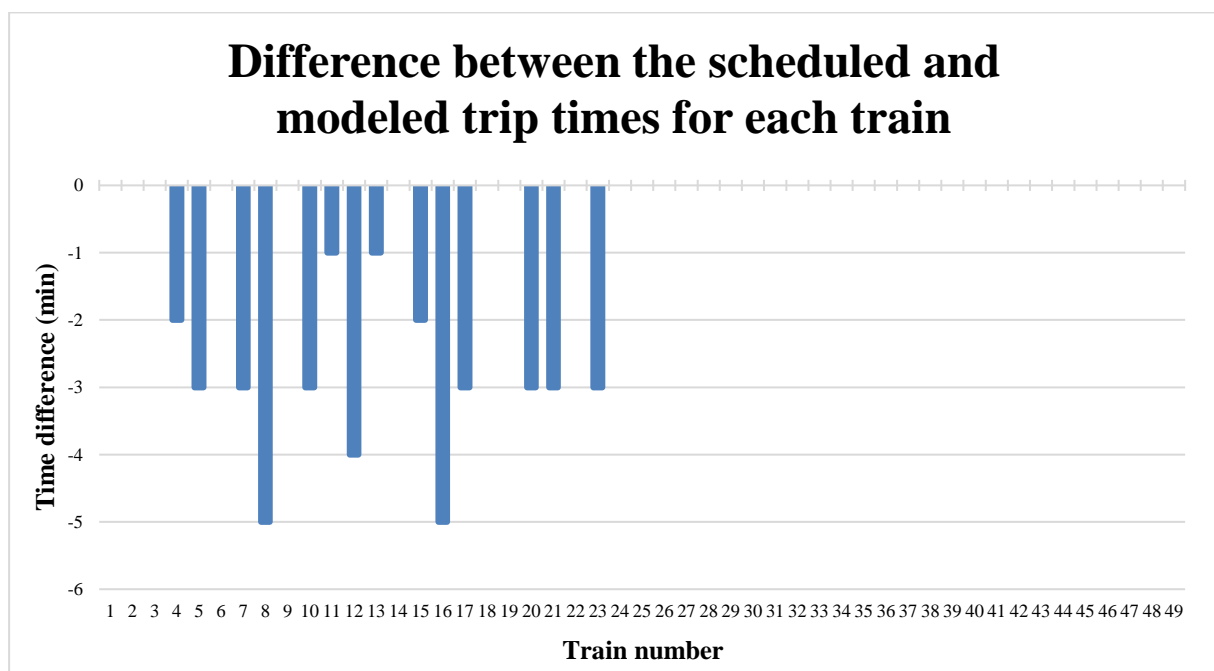


Figure 5-7: The difference between the scheduled and modelled trip times when the model is run without delays⁷

5.4 Validation with delays

After the speed properties of the trains are calibrated, it is also necessary to calibrate the amount of delays so that the trains' effective delays match that of reality. As explained in Section 4.5.1, simply instigating the real amount of delays, will not necessarily cause the real duration of delays in the model. In this Section, the calibration of the delays will be discussed. This will also be the last step to the validation of the model. If the model is found to be valid and representative of reality, new trains will be added to the model to finally answer the research questions asked in Chapter 1.

⁷ The model was calibrated so that the majority of trains have exactly the same trip time as specified in the schedule. There are however 14 trains in the schedule that have shorter trip times for the same route and therefore the corresponding trains in the model have different trip times. This difference in trip time is usually to accommodate other trains from other routes that share the same track in some areas of the line.

To determine if the model is calibrated the following three output parameters will be compared to what was seen in data for a 6 month period:

- Number of delays
- Total delay minutes
- Average delay duration

The input parameter that will be adjusted is the primary delay frequency. The delay frequencies for both peak and off-peak periods will be increased to compensate for the limitations of the model explained in Sections 4.5.1 and 4.5.4. Because of the randomness of the model, it is necessary to run the simulation multiple times for every change of delay frequency to ensure that the model output deviations converge to a reasonable number. It was found during the initial calibration rounds that after 5 runs of the same calibration setting, that the model produced the outputs of all three parameters within a standard deviation 5%. Therefore, for the final calibration rounds the model was run 5 times. Table 5-4 shows the results of the last calibration round⁸. Note that the standard deviation values in column 3 of Table 5-4 indicate the standard deviation of the delay durations for each run. The standard deviations at the bottom of the table are relevant to the average of all the runs.

Table 5-4: Last calibration round results

1/04/2015 - 30/09/2015				
Run	Mean delay duration	Std dev	Total minutes delay	Number of delays
1	0:17:25	0:12:05	63666	3656
2	0:16:51	0:12:12	61124	3629
3	0:16:35	0:11:29	60797	3668
4	0:17:21	0:10:28	65486	3775
5	0:17:37	0:11:47	65213	3703
Average	0:17:09	0:11:36	63257	3686
Std dev	0:00:23	0:00:37	1978	50
Std dev %	2%	5%	3%	1%

The primary delay frequency that resulted in these outputs was 118 min and from Table 5-2 we see that the real delay frequency was 131 min (10% difference). Because 42% of the time entries were faulty as discussed in Section 5.2.1, it was not possible to calculate separate frequencies for peak and off-peak periods from data. This shows that when the model is calibrated (Total delay minutes parameter is close enough to the observed value from data), the time between delays in the model is 10% shorter than what data shows. The model frequency however is expected to be higher since the model can only simulate one delay at a time, whereas in reality delays can happen concurrently. It must also be considered that the real deal frequency was calculated by dividing the total operating time over six months by total number of delays. If for example three delays of different durations occurred

⁸ Each run shown in Table 5-4 represents a simulation of 6 months with no new trains (i.e. the base case). The Mean delay duration is therefore the average of all the delays that occurred in the simulated 6 months. The same applies for the standard deviation. The Total minutes delay and Number of delays are summation parameters of all the delays that was simulated in the modelled 6 months.

in a time period of ten minutes, the delay frequency would be 3.33min, regardless of the real time between delays and regardless of the possibility of occurring concurrently. The frequency of delays in the model however was a pre-set fixed value, meaning that if three delays occurred in the 10min period, the time between them will always be 2 min and therefore the frequency 2min. The model can therefore simulate the same amount of delays with a higher frequency.

Table 5-5 compares the results of the final calibration run shown in Table 5-4 to that calculated from data for both primary and secondary delays. Note that the total number of delays and total minutes in the “All delays” column of Table 5-4 is half of the number of delays in Table 5-5. This is because data shows delays in both directions of the line, while only one direction was modelled. The number of delays and total minutes modelled are thus doubled in order to be able to compare the modelled delays with the actual delays.

In the “Primary delays” column, delay duration and total minutes delay have large differences when compared to the values extracted from data. Number of delays, however match very well with the difference being less than a percent. When the “All delays” column is studied it can be noticed that total minutes delay now match with only a 3% difference, but number of delays and delay duration differ with -19.7% and 28.5% respectively. From the number of delays it appears that the simulation modelled approximately 23.5% less secondary delays (*All delays – Primary delays*) than what occurred in reality. This means that either the simulation modelled the effect of primary delays on secondary delays incorrectly or that the assumption made concerning which delay entry indicates the primary delay, was wrong (refer to Section 5.2.3).

Table 5-5: Summary of primary delays and resulting sum of delays modelled compared to data

	Primary delays			All delays		
	Data	Model	Difference	Data	Model	Difference
Average delay duration	00:14:24	00:17:19	20.3%	00:13:21	00:17:09	28.5%
Standard deviation	00:14:00	00:11:36	-17.1%	00:14:18	00:11:36	-18.9%
Number of delays	1424	1429	0.3%	9188	7372	-19.7%
Total minutes	20448	24750	21%	122789	126514	3.0%

The only way the model could be wrong is if the trips times are inaccurate or if the schedule was incorrect. Figure 5-7 shows that 71% of the modelled trains’ trip times are exactly the same as on the schedule. The other 29% of trip times are between 1- and 5 minutes off the schedule. These differences are too small to have caused such a significant error in number of delays. Delay duration would not have been affected by trip time. Furthermore, the schedule that was used in the model was identical to the schedule that was in operation at the time of data capture.

It can therefore be concluded that the assumption that the first entry of a delay log sheet is the primary delay is wrong. This statement can further be substantiated by considering Figure 5-8 which is a plot of all the assumed primary delays and the resulting sum of the delays. It is clear that there is no

relationship between primary delays and the subsequent sum of delays. Calibrating the model by use of comparison with primary delays is therefore impossible.

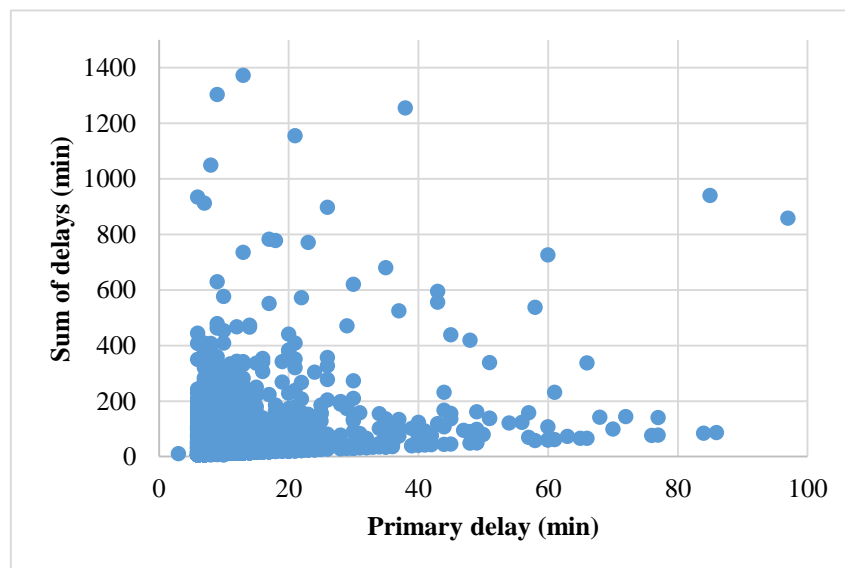


Figure 5-8: Scatter plot of the relationship between primary delays and the resulting sum of delays

In the “All delays” column in Table 5-5 the differences between the model outputs and data are shown. To determine if these differences are too large for the model to represent reality sufficiently (i.e. if the model is valid), statistical methods were used to test the validity of the model. The two most common procedures used in simulation modelling for validation testing are the Chi-square and Kolmogorov-Smirnoff (K-S) goodness-of-fit tests [44]. In the case of this model the K-S goodness-of-fit test procedure was used to determine if the modelled data sufficiently represents the observed data. Since the output data of this model is continuous, Kelton & Law [44] states that both Chi-square and K-S tests can be used, however the Chi-square test is very dependent on the sample size and the way data is binned (frequency bins). Since the sample size of the observed data was limited to 6 months (26 weeks or 26 samples) it was opted to use the K-S test.

The K-S test compares the largest vertical difference between the observed and estimated (in this case the modelled) cumulative distributions at any point to that of a critical value that depends on the level of confidence required [44]. The critical value is read from a table, which in this case was provided by Law & Kelton [44], that summarises the critical values for the K-S goodness-of-fit test. The following Sections will describe how the K-S procedure was followed to determine the validity of the model in terms of the Number of delays, Total delay minutes and Average delay duration parameters. Table 5-6 shows the values that will be used for the Null Hypothesis of each parameter. These were calculated from all the delays observed in data.

Table 5-6: Mean and variance values for each parameter calculated from the observed data

	Observed data		
	Total weekly minutes delay	Number of weekly delays	Delay duration
μ	4703	351	13
σ^2	3154810	14162	204

5.4.1 Number of delays

To validate the output of the model the three parameters have to be analysed separately. Since the modelled schedule is repeatable every week in terms of number of trains and departure times; Number of delays and Total minutes delay were sampled after each modelled week. Modelling six months therefore provides 26 samples that can be tested against the 26 samples from the observed data. This Section will elaborate on the K-S test procedure to determine the validity of the Number of delays.

1 Null Hypothesis – H_0

The observed data have a theoretical cumulative distribution $F(X)$ with mean $\mu = 351$ and $\sigma^2 = 14161$.

2 Calculation of right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n

$$D_n^+ = \max_{0 \leq i \leq n} \{F_n(X_i) - F(X_i)\} \quad (4)$$

$$D_n^- = \max_{0 \leq i \leq n} \{F(X_i) - F_n(X_{i-1})\} \quad (5)$$

$$D_n = \max\{D_n^+, D_n^-\} \quad (6)$$

Table 5-7: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for Number of delays per week

Number of delays per week						
		Observed		Model Estimate		K-S statistic
	X_i	Frequency	$F(x)$	Frequency	$F_n(x)$	D_n
X_1	0	0	0	0	0	0
X_2	50	0	0	2	0.01	0.01
X_3	100	0	0	2	0.03	0.03
X_4	150	0	0	0	0.03	0.03
X_5	200	1	0.04	0	0.03	0.01
X_6	250	4	0.19	11	0.11	0.08
X_7	300	5	0.38	98	0.84	0.46
X_8	350	5	0.58	21	1	0.42
X_9	400	6	0.81	0	1	0.19
X_{10}	450	1	0.85	0	1	0.15
X_{11}	500	1	0.88	0	1	0.12
X_{12}	550	1	0.92	0	1	0.08
X_{13}	600	0	0.92	0	1	0.08
X_{14}	650	1	0.96	0	1	0.04

Number of delays per week					
		Observed		Model Estimate	
	X_i	Frequency	$F(x)$	Frequency	$F_n(x)$
X_{15}	700	1	1	0	1
					D_n
					0.00

3 K-S goodness-of-fit test

As seen in Table 5-7 the supreme value (coloured in red) for D_n is 0.46. According to Law & Kelton[44] to compare D_n to the critical value - $c_{1-\alpha}$ - the following test condition applies:

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n > c_{1-\alpha} \quad (7)$$

For $n = 26$, $1 - \alpha = 0.975$ and $c = \mathbf{1.48}$ the result is:

$$2.40 > \mathbf{1.48}$$

Therefore the H_0 is rejected, and the Number of delays parameter produced by the model cannot be deemed valid. The observed and modelled cumulative distributions can be seen in Figure 5-9. It is clear that the model under-estimates the number of delays per week. The modelled distribution is also much narrower than the observed data. This can be explained by the fixed frequency by which delays are instigated, resulting in similar amounts of delays happening in each week. However, the reason why the modelled distribution shows some variation in the amount of delays is because the time between delays are dependent on both the previous delay duration (which is a random amount) and the fixed frequency as explained in Section 4.3.

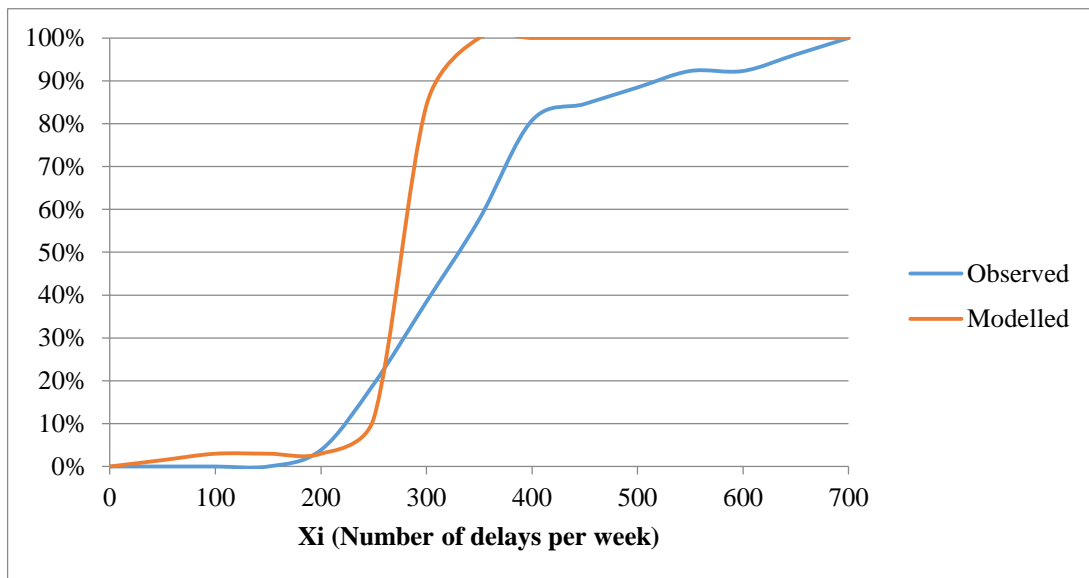


Figure 5-9: Cumulative distributions describing the Number of delays per week from the observed and modelled data sets

5.4.2 Total minutes delay

As mentioned in Section 5.4.1, Total minutes delay will also be tested by sampling the minutes delayed in each week giving a sample size of 26. The same K-S procedure will be followed as with Number of delays.

1 Null Hypothesis – H_0

The observed data have a theoretical cumulative distribution $F(X)$ with mean $\mu = 4703$ and $\sigma^2 = 3154810$.

2 Calculation of right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n

Table 5-8: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for Total minutes delayed per week

Total minutes delay per week						
		Observed		Model Estimate		K-S statistic
	X_i	Frequency	$F(X)$	Frequency	$F_n(X)$	D_n
X1	0	0	0	0	0	0
X2	500	0	0	1	0.01	0.01
X3	1000	0	0	3	0.03	0.03
X4	1500	0	0	0	0.03	0.03
X5	2000	0	0	0	0.03	0.03
X6	2500	2	0.08	0	0.03	-0.05
X7	3000	4	0.23	0	0.03	-0.20
X8	3500	1	0.27	6	0.07	-0.19
X9	4000	3	0.38	28	0.28	-0.10
X10	4500	3	0.50	36	0.55	0.05
X11	5000	1	0.54	34	0.81	0.27
X12	5500	4	0.69	16	0.93	0.23
X13	6000	3	0.81	8	0.99	0.18
X14	6500	3	0.92	1	0.99	0.07
X15	7000	0	0.92	1	1	0.08
X16	7500	0	0.92	0	1	0.08
X17	8000	0	0.92	0	1	0.08
X18	8500	0	0.92	0	1	0.08
X19	9000	2	1	0	1	0

3 K-S goodness-of-fit test

As seen in Table 5-8 the supreme value (coloured in red) for D_n is 0.27. According to Law & Kelton[44] to compare D_n to the critical value - $c_{1-\alpha}$ - the following test condition applies:

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n > c_{1-\alpha} \quad (7)$$

For $n = 26$, $1 - \alpha = 0.975$ and $c = 1.48$ the result is:

$$1.40 < 1.48$$

Therefore, there is not enough evidence to reject H_0 . The model can therefore be considered valid for the Total delay minutes parameter. From Figure 5-10 it can be seen that again the model produced a narrower distribution than what is seen from data. The reason is that the model did not include the major outliers from data into the delay duration distributions used to determine the length of delays. Including the outliers would cause the model to be excessively random. If the model is too random, it becomes difficult to produce results with a small enough variance to be able to explain the behaviour of a system. The goodness-of-fit test however indicates that the model's estimation of total weekly delay minutes is close enough to be able to consider the model valid for this parameter.

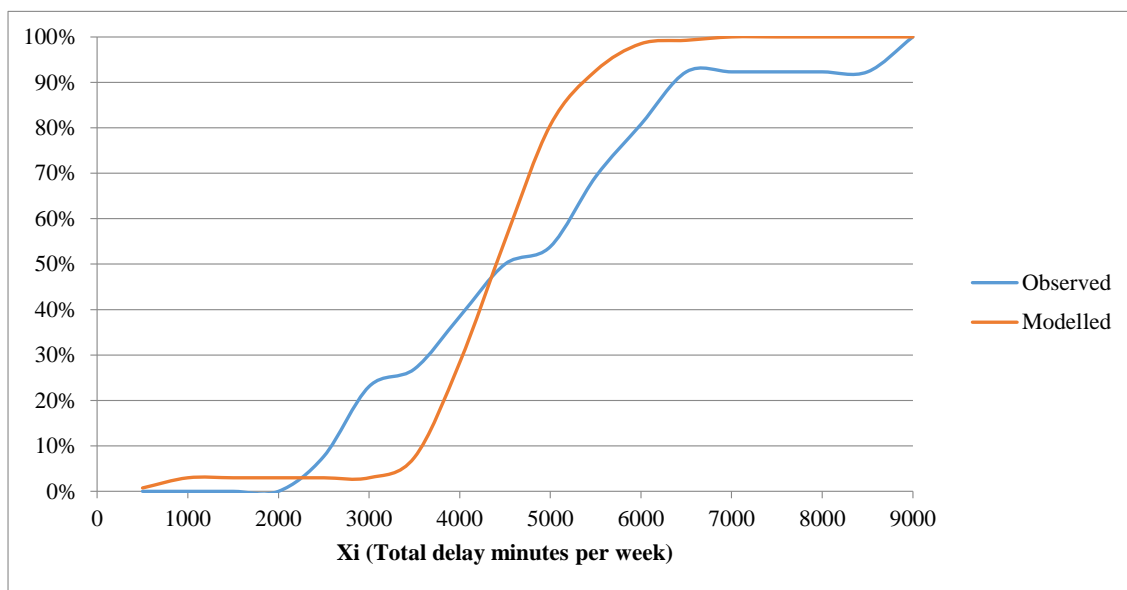


Figure 5-10: Cumulative distributions describing the Total minutes delay per week from the observed and modelled data sets

5.4.3 Average delay duration

To validate the Average delay duration parameter the same approach was followed as with Number of delays and Total minutes delayed parameters, where the Average delay duration was calculated after each week. The value of Average delay duration is therefore simply a function of Number of delays and Total delayed minutes.

1 Null Hypothesis – H_0

The observed data have a theoretical cumulative distribution $F(X)$ with mean $\mu = 13$ and $\sigma^2 = 204$.

2 Calculation of right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n

Table 5-9: Summary of the right-continuous step-function $F(X)$, $F_n(X)$ and K-S statistic D_n for Average delay duration per week

Average delay duration per week						
		Observed		Model Estimate		K-S statistic
	X_i	Frequency	$F(X)$	Frequency	$F_n(X)$	D_n
X1	0	0	0	0	0	0
X2	8	0	0	0	0	0
X3	9	0	0	1	0.01	-0.01
X4	10	2	0.08	0	0.01	0.07
X5	11	2	0.15	0	0.01	0.15
X6	12	5	0.35	0	0.01	0.34
X7	13	6	0.58	2	0.02	0.55
X8	14	3	0.69	9	0.09	0.60
X9	15	1	0.73	29	0.31	0.42
X10	16	1	0.77	31	0.54	0.23
X11	17	3	0.88	33	0.78	0.10
X12	18	1	0.92	18	0.92	0.01
X13	19	2	1	6	0.96	0.04
X14	20	0	1	5	1	0

3 K-S goodness-of-fit test

As seen in Table 5-9 the supreme value (coloured in red) for D_n is 0.60. According to Law & Kelton[44] to compare D_n to the critical value - $c_{1-\alpha}$ - the following test condition applies:

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n > c_{1-\alpha} \quad (7)$$

For $n = 26$, $1 - \alpha = 0.975$ and $c = 1.48$ the result is:

$$3.16 > 1.48$$

H_0 is therefore rejected and it can be concluded that the model is not valid for the Average delay duration parameter. From Figure 5-11 it can be seen that the model over-estimates the Average delay duration.

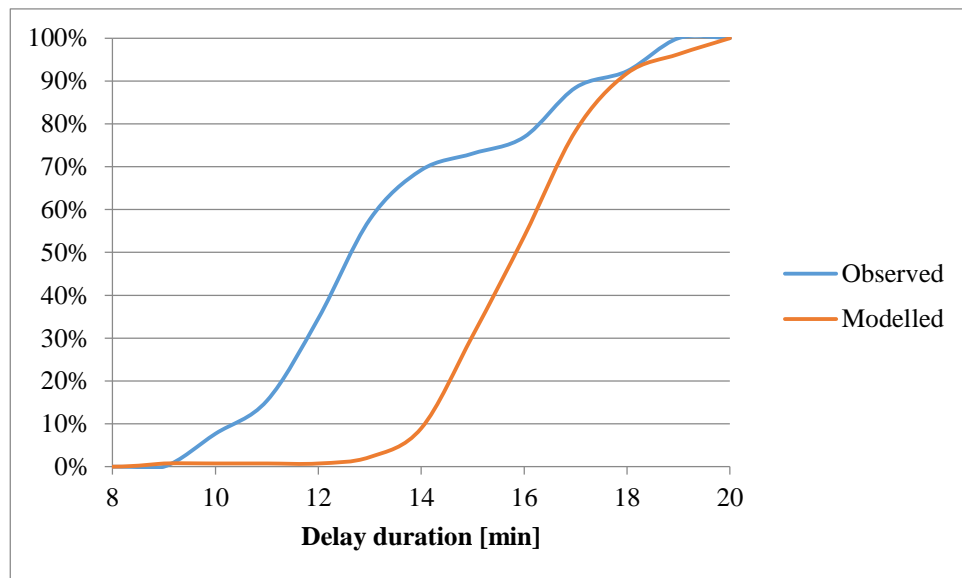


Figure 5-11: Cumulative distributions describing the Average delay duration per week from the observed and modelled data sets

5.4.4 Conclusion

It can be concluded that only one of the three test parameters for this model is valid according to the K-S goodness-of-fit test. Figure 5-9 shows that the model under-estimates the Number of delays per week, while Figure 5-11 shows that the model over-estimates Average delay duration per week. For the Total delay minutes parameter, however it was found that the model's cumulative distribution is slightly narrower than the observed distribution (reference to Figure 5-10) and that the K-S test found the model's distribution a good fit.

Since only one parameter is found to give a good fit to the observed data, the usefulness of the model can be questioned. However, as explained in Section 4.1.1 the key parameter by which punctuality is measured is Total minutes delayed which in this case was found to be the valid parameter. It was found that the way the model is constructed and the assumptions that had to be made, forces the modeller to compromise the validity of Number of delays and Average delay duration parameters in order to ensure the validity and distribution fit of the key punctuality parameter - Total minutes delayed.

Even though not all the parameters were found representative of reality, it must be understood that the same model will be used to compare different scenarios in the same simulation environment. Additionally, the size of the fault in Number of delays and Average delay duration can be expected to be consistent since the standard deviations of 70 model runs (as mentioned in Section 5.4) for these two parameters are within 5%.

6. Scenarios and model outputs

6.1 Overview

PRASA's plan is to phase in the new trains as they are rolled out of production because of various political and operational constraints. There are several factors that have to be taken into account as to where these new rolling stock are commissioned. It has to be mentioned that socio-economic factors play a major role in projects like this, especially in a country like South Africa. Thus besides the technical complexities that have to be considered, there also exists a social factor that might enjoy priority above the technical factors. Socio-economic factors do not fall in the scope of this study.

The fleet currently running on the Chris Hani to Cape Town route has 14 old trains shuttling to and from Cape Town. The model was run first to be calibrated with no new trains in the fleet as discussed in previous chapters. Thereafter one new train was added to the fleet replacing one old train for each scenario. New trains were added until all 14 old trains were replaced with new trains. As shown in the calibration stage of the model, each Scenario had to be run 5 times in order to get a reasonable standard deviation for each of the performance measures. This will be the base case.

The subsequent case will simulate the same scenarios but with an improvement of the system reliability as a whole by 50%. This Case is included to test the effect on punctuality if not only rolling stock related delay causes are mitigated. Fifty percent is simply an arbitrary value to create a different simulation environment. In practice this implies that all the components that have the potential to cause a train delay will be assumed to have been refurbished or maintained to such an extent that they have a 50% less chance of experiencing a delay-causing failure. In terms of the model, this was implemented by changing the *systemRel* factor to $(1 + 0.5)$ resulting in the delay frequency to be lengthened by 50% (refer to Figure 4-7).

This chapter will therefore cover the simulation results of the following 2 cases:

1. Case 1 – 0% system improvement but with the incremental addition of new trains that will eventually mitigate rolling stock related delays.
2. Case 2 – 50% system improvement and with the incremental addition of new trains that will eventually mitigate rolling stock related delays.

It is assumed that the new trains are 100% reliable and therefore will not experience any rolling stock related delays. For Case 1 it is then expected that after all 14 old trains are replaced with new trains the 25% contribution of rolling stock related delays to the total minutes delay, as seen in Figure 6-2, will be mitigated and the total will be approximately 25% less than with zero new trains. The same applies to the number of delays, where rolling stock contributes 26% (see Figure 6-1).

For Case 2 it is expected that for all the scenarios, the number of delays and total minutes delay will be 33% less than for Case 1.

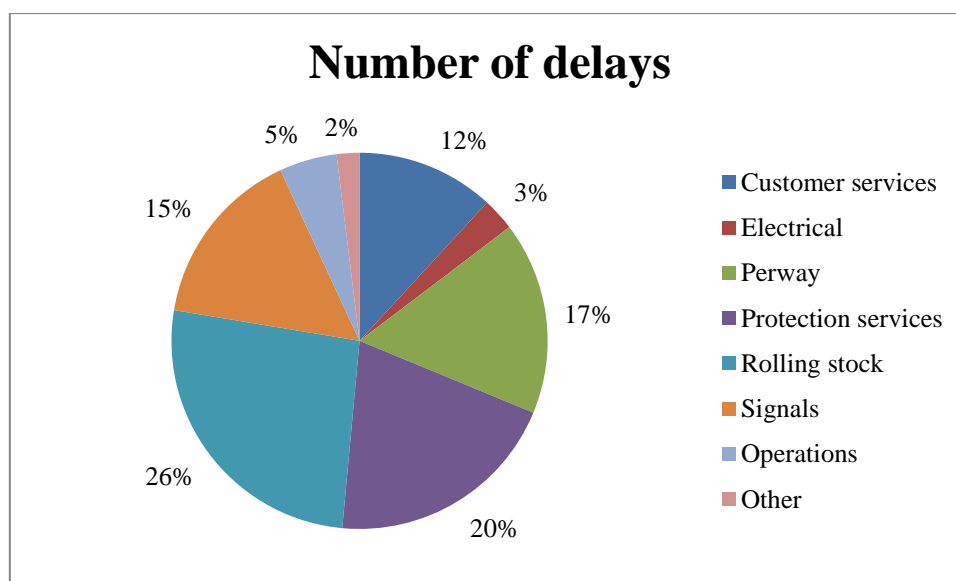


Figure 6-1: Number of delays for each department

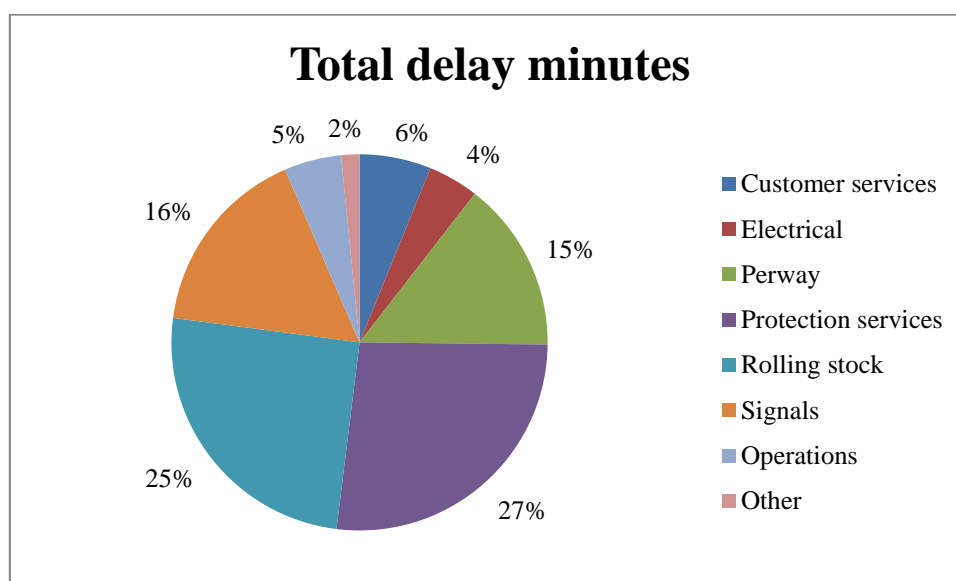


Figure 6-2: Total delay minutes for each department

6.2 Case 1

Figure 6-3 shows number of delays for each scenario. Here we see a 29% decrease in number of delays which is very close to the expected the 26% (refer to Figure 6-1) which refers to the percentage of delays related to rolling stock. Since the new trains are 100% reliable, replacing the whole fleet will mean that no more rolling stock related delays will occur. The extra 3% decrease predicted by the model accounts to 110 delays over the 6 months simulated period which means

that the model simulated one delay too little every 2.4 days. The 3% difference can therefore be accounted to the stochastic method with which delays were instigated.

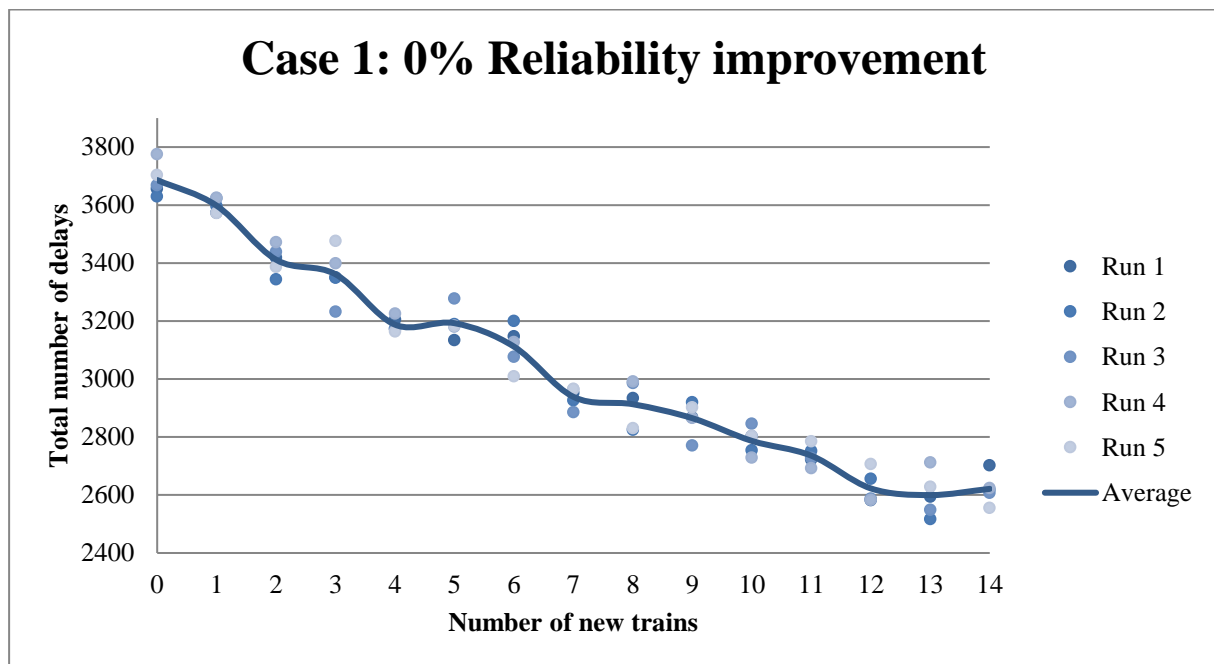


Figure 6-3: Number of delays for Scenarios 0-14 and Case 1

In Figure 6-4 a 37% decrease in total minutes delayed is observed. This is significantly more than the expected 25%. This can be explained by the 11% decrease in average delay duration shown in Figure 6-5, i.e. even though the number of delays decreased by the predicted 29%, the additional decrease in duration of delays resulted in the total minutes delay to decrease by a further 11%.

In practical terms this can be explained by imagining the journey of two consecutive trains. In Scenario 0, train B would run and be delayed for 5min (secondary delay) by train A ahead busy experiencing a rolling stock delay which started 5min earlier (i.e. 10min primary delay). When train B then proceeds again, it will run and later experience its own primary delay of 12min related to a perway failure. If train B then arrives at Cape Town station the model will calculate train B experienced one delay of 17min (5min + 12min). For both train A and B the model calculates two delays with a total of 27min at an average delay duration of 13.5min.

In Scenario 14 for the same situation, train A will be a new train and will therefore not experience the 10min primary delay. Train B (also a new train) will then only experience the perway related delay of 12min and in the end the model will calculate one delay of 12min at an average delay duration of 12min for both train A and B.

It is now clear that by introducing new rolling stock the number of delays were reduced by a predictable one delay, but the total delay minutes and average delay duration was reduced by less predictable amounts of 15min and 1.5min respectively.

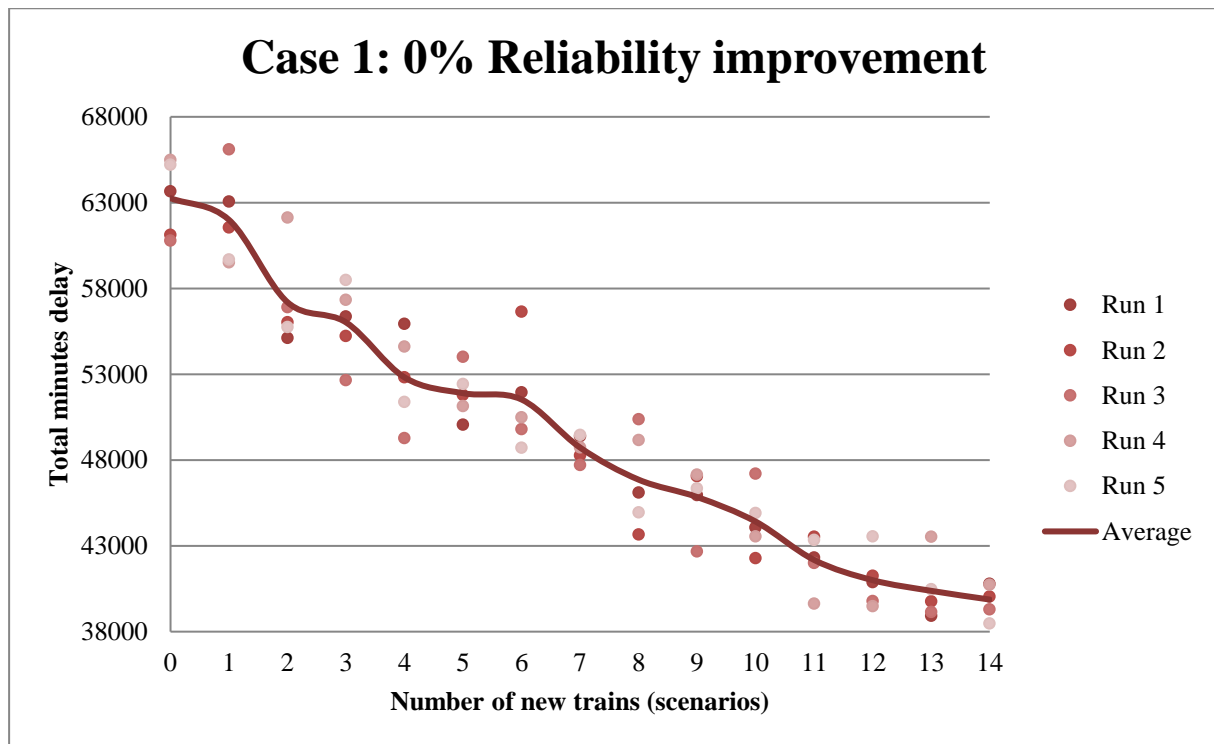


Figure 6-4: Total sum of delays for Scenarios 0-14 and Case 1

Figure 6-5 shows how the mean delay duration changed for each scenario or the adding of new trains. An 11% improvement can be seen from 0 new trains to 14 new trains. The average line might seem rather un-even, however the large increase seen from 5 new trains to 7 new trains is but 21 seconds (2%). This un-expected deviation is thus relatively small, and can be accounted to the random nature of the model. The mean delay duration decreased because the new trains are assumed to be 100% reliable which leads to an incremental reduction of rolling stock related primary delays. Rolling stock was found to have the largest average delay duration of all the departments. Therefore by eliminating rolling stock related delays it can be expected that the mean delay duration will decrease. The 1% extra decrease in delays predicted by the model can also be accounted to the stochastic nature of the model. This figure is expected to move closer to 26% if more simulations runs are done.

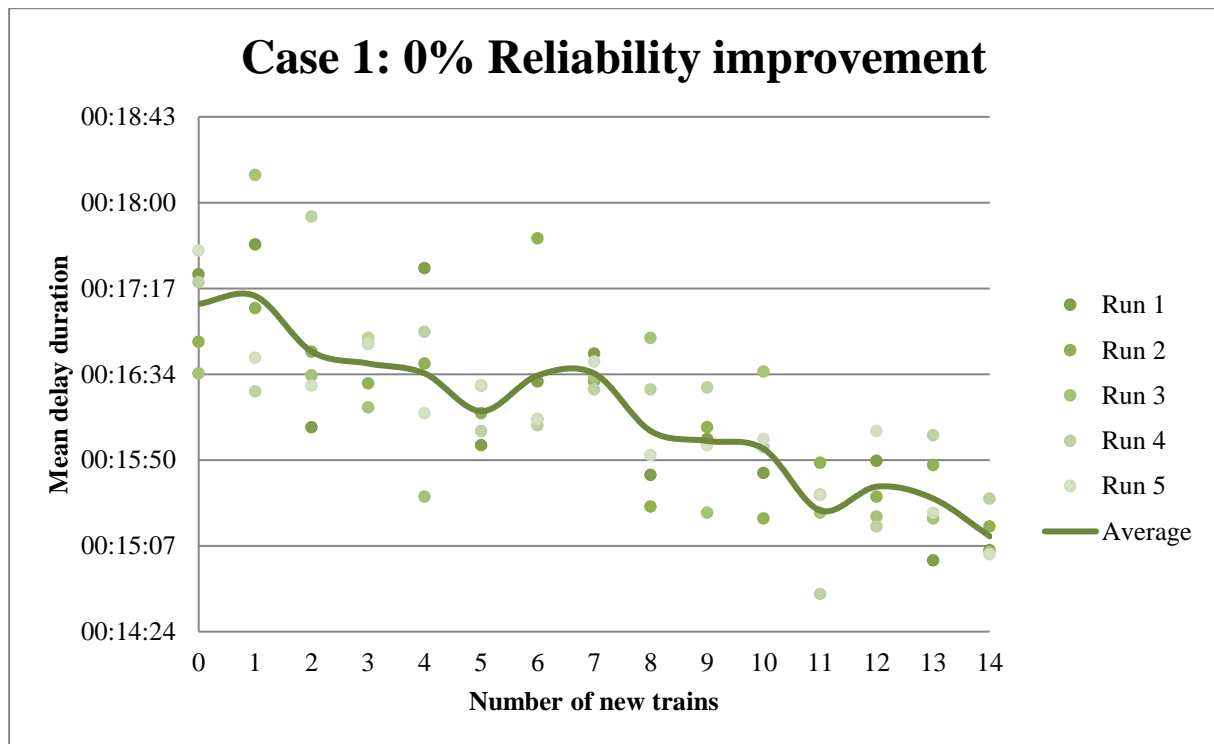


Figure 6-5: Mean delay duration for Scenarios 0-14 and Case 1

6.3 Case 2

As mentioned, in Case 2 the system was assumed to be 50% more reliable in all the departments. This was facilitated by simply lengthening the frequency of delays by 50%.

In Figure 6-6 an overall decrease of 31% can be seen between Scenario 0 and 14. This is a 2% larger overall decrease than what was seen in Case 1. A more detailed comparison of the two cases is covered in Section 6.4.

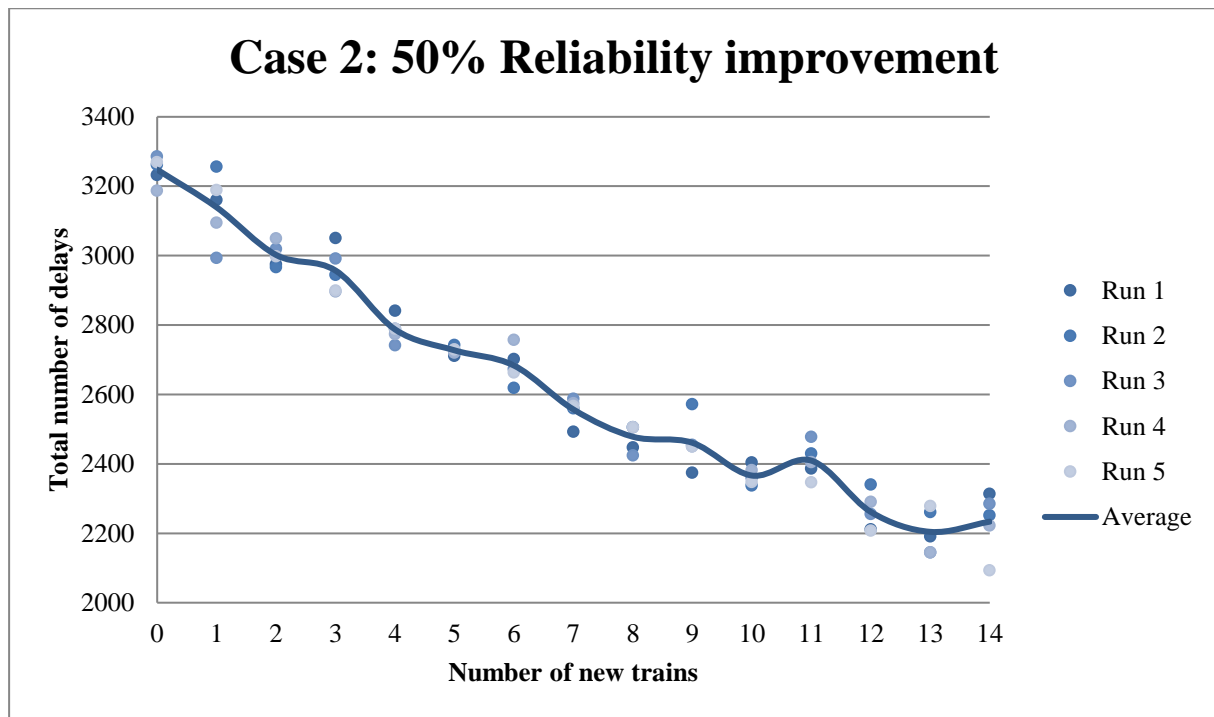


Figure 6-6: Number of delays for Scenarios 0-14 and Case 2

The total minutes delay for Case 2 can be seen in Figure 6-7. Similar to the number of delays, a larger overall decrease of 36% in Case 2 is observed than what is observed in Case 1. This is because, as also explained in Section 6.2, when less primary delays occur; secondary delays reduce by a larger amount.

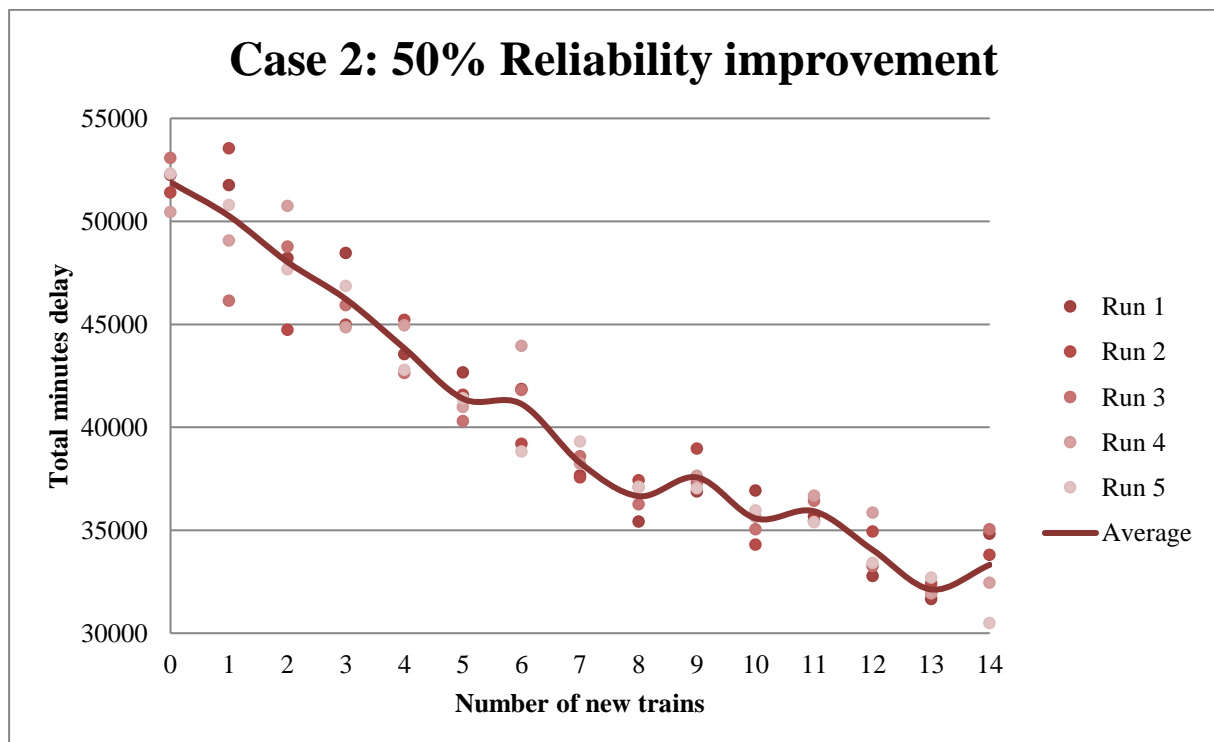


Figure 6-7: Total minutes delays for Scenarios 0-14 and Case 2

An exponential relationship between primary delays and sum of delays was observed by Hwang & Liu [25] (refer to Section 3.2.2), however in this model a parabolic relationship was observed as seen in Figure 6-8. Hwang & Liu's [25] model was different in that they modelled a line to test the effect on

sum of delays by varying the length of a single primary delay for a specific time of day. An exponential relationship between the duration of a primary delay and the sum of delays is therefore an expected result. In Figure 6-8 the sum of all the primary delays and total delays over a period of six months are plotted, and therefore the same of the curve differs. Nevertheless the insight to be acquired from Figure 6-8 is that the relationship between primary delay duration and the sum of delays is not linear and therefore if primary delays are reduced an extended amount of total delay minutes can be saved.

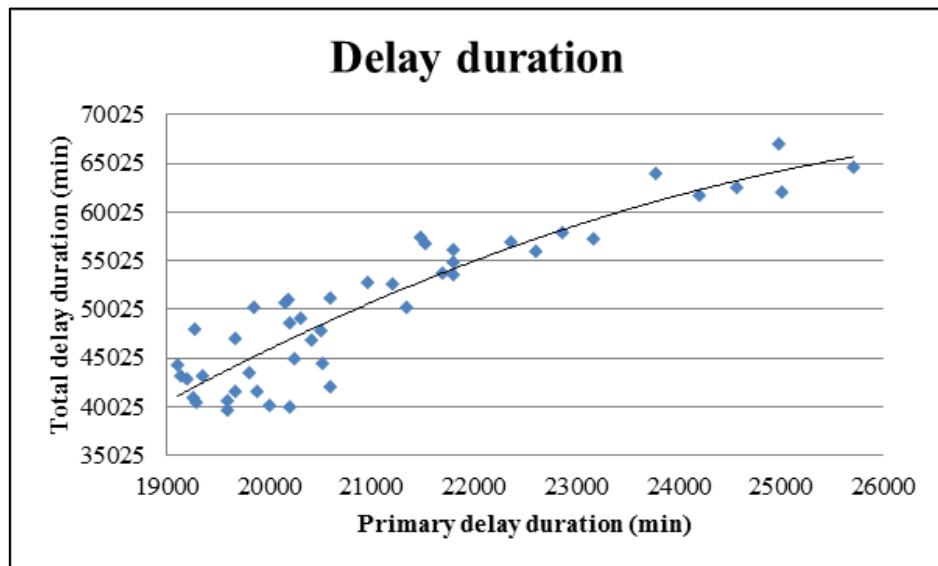


Figure 6-8: Relationship between primary delay duration and sum of delays from modelled data

Figure 6-9 shows a 7% decrease in mean delay duration when comparing Scenario 0 and 14.

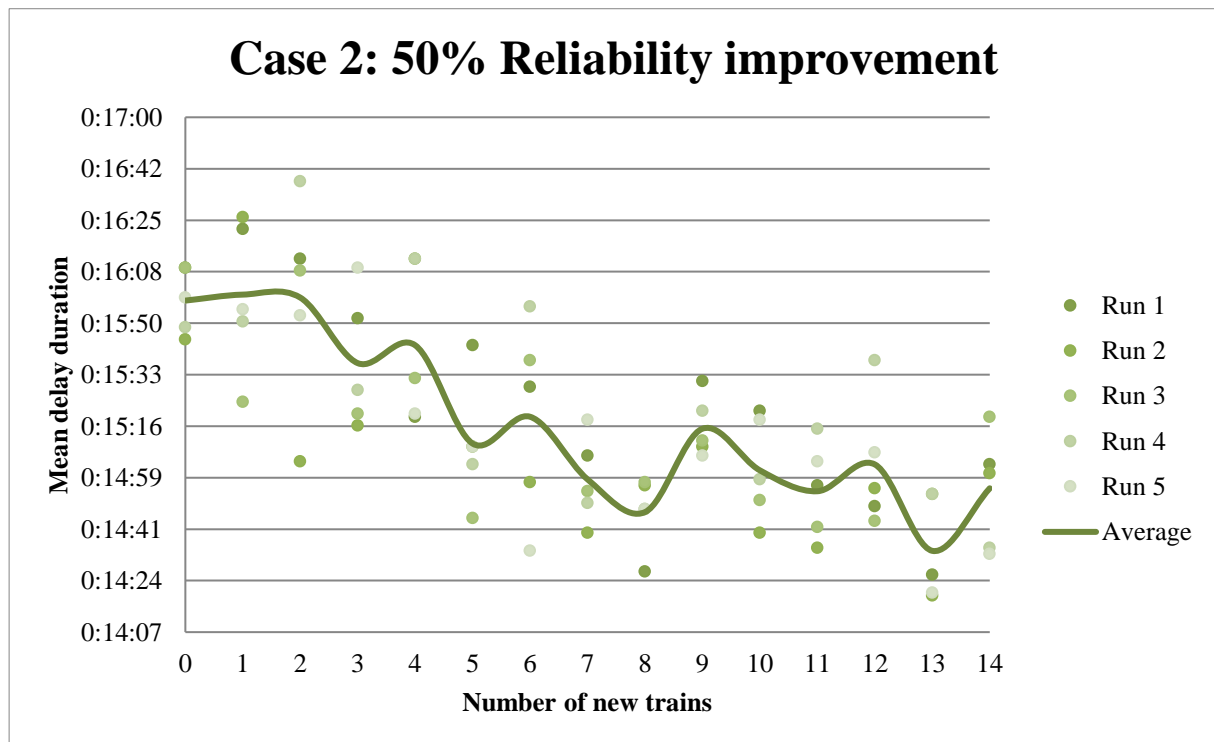


Figure 6-9: Mean delay duration for Scenarios 0-14 and Case 2

6.4 Comparison of Case 1 and Case 2

In this Section the results for Case 1 and Case 2 will be compared. This will provide insights to how the addition of new trains and improvement in system reliability will affect the three performance parameters used to measure punctuality namely, delay duration, number of delays and total minutes delay. A summary of the results discussed in this section is shown in Table 6-1.

6.4.1 Number of delays

The number of delays improved by an average of 13% across all the Scenarios. In Scenario 0, Case 2 is 12% (438 delays) less than Case 1 and in Scenario 14, Case 2 is 15% (388 delays) less than Case 1 (see Figure 6-10). The total improvement from Case 1, Scenario 0 to Case 2 Scenario 14 is 39% (1453 delays). Since the frequency of delays was lengthened by 50% in Case 2, it is expected that for the same scenario, there would be 33% less delays in Case 2. However, following the same argument made in Section 6.2, it can be explained that if a train in a 0% improved environment, experiences 2 delays during a trip, it will only be counted as one delay when it arrives at the destination station. In a 50% improved environment the same train will for instance only experience 1 delay during its trip, which then will also only be counted as one delay at the destination station. The effect of the reliability improvement is therefore diluted in the measurement of the number of delays.

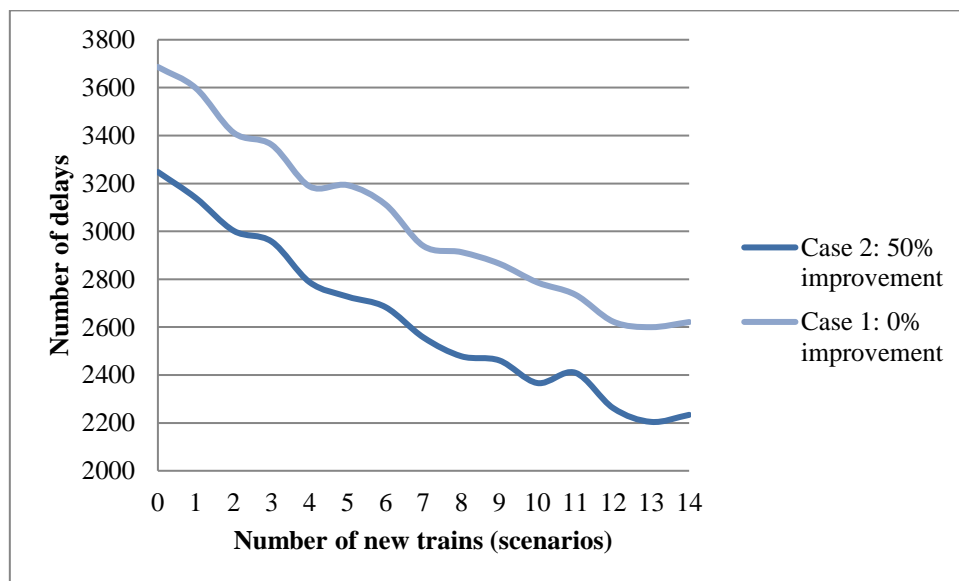


Figure 6-10: Number of delays comparison of Case 1 and Case 2

6.4.2 Total minutes delay

The total minutes delay improved by an average of 19% across all scenarios. In Scenario 0, Case 2 is 18% (11 360 min) less than Case 1 and in Scenario 14, Case 2 is 16% (6 548 min) less than Case 1 (see Figure 6-11). The total improvement from Case 1, Scenario 0 to Case 2 Scenario 14 is 47% (29 934 min). Since the frequency of primary delays are lengthened and new rolling stock is introduced, resulting in an even larger decrease in primary delays, the same extended reduction in total delay minutes is seen as when Case 1 and 2 was studied individually.

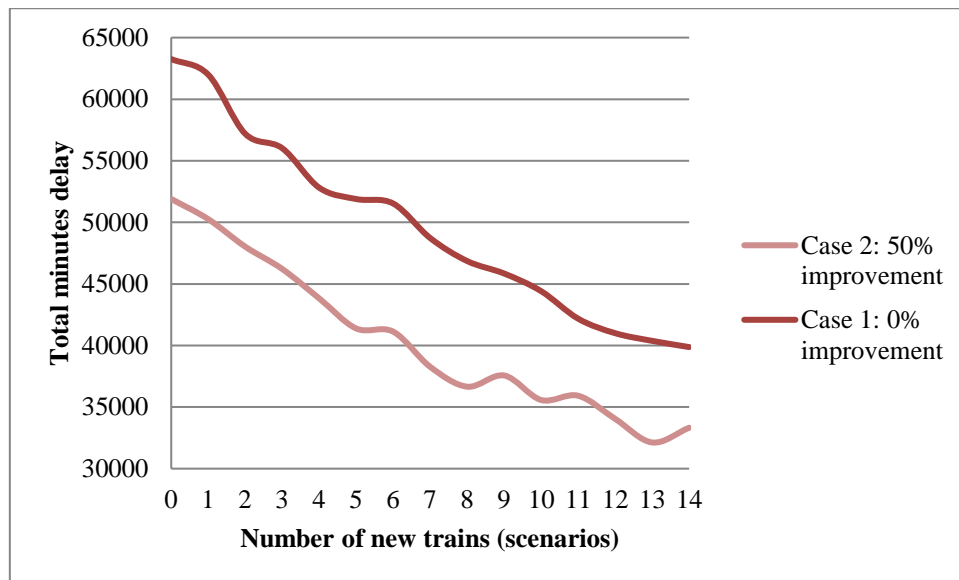


Figure 6-11: Total minutes delay comparison of Case 1 and Case

6.4.3 Mean delay duration

The mean delay duration improved by an average of 6% across all Scenarios. Improvements of 7% for Scenario 0 and 2% for Scenario 14 (see Figure 6-12) were also made. The total improvement from Case 1, Scenario 0 to Case 2, Scenario 14 is 13% (0:02:15). Since reliability improvement from Case 1 to Case 2 was enforced by means of only lengthening the time between primary delays, it can be questioned why mean delay duration decreases at all. The same concept explained in Section 6.2 applies where it is possible for a train to experience more than one delay per trip. In a 0% improved environment it is more likely for a train to experience two or three delays per trip than in a 50% improved environment. Furthermore, as mentioned, in this model a train's delay duration is measured at the destination station (i.e. the sum of the train's delays for that trip), and therefore if the number of delays per trip reduce with system improvement the mean delay duration will also decrease.



Figure 6-12: Mean delay duration comparison of Case 1 and Case 2

Table 6-1: Summary of results

Scenario	Number of delays			Total delay minutes			Average delay duration		
	Case 1	Case 2	% Change	Case 1	Case 2	% Change	Case 1	Case 2	% Change
0	3686	3248	12%	63257	51897	18%	0:17:09	0:15:59	7%
1	3598	3139	13%	61993	50259	19%	0:17:14	0:16:00	7%
2	3411	3002	12%	57192	48025	16%	0:16:46	0:16:00	5%
3	3361	2957	12%	56019	46214	18%	0:16:40	0:15:38	6%
4	3187	2787	13%	52813	43824	17%	0:16:34	0:15:43	5%
5	3193	2727	15%	51894	41383	20%	0:16:15	0:15:10	7%
6	3112	2683	14%	51526	41125	20%	0:16:33	0:15:19	7%
7	2938	2557	13%	48724	38269	21%	0:16:35	0:14:58	10%
8	2913	2478	15%	46855	36656	22%	0:16:04	0:14:47	8%
9	2865	2461	14%	45844	37565	18%	0:16:00	0:15:16	5%
10	2787	2366	15%	44411	35575	20%	0:15:56	0:15:02	6%
11	2736	2409	12%	42176	35921	15%	0:15:25	0:14:55	3%
12	2623	2262	14%	40999	34044	17%	0:15:38	0:15:03	4%
13	2600	2204	15%	40377	32127	20%	0:15:32	0:14:35	6%
14	2621	2234	15%	39871	33323	16%	0:15:13	0:14:55	2%
Average improvement			13%			19%			6%
Case improvement	29%	31%		37%	36%		11%	7%	
Overall improvement		39%			47%			13%	

7. Conclusion

The objective of the study was to determine what the effect will be on punctuality when adding new trains to a fleet of old trains. The following three performance parameters were used to measure punctuality:

- Total number of delays
- Total minutes delay
- Mean delay duration

A case study was made of the Western Cape Metrorail network with specific focus on the line between Chris Hani and Cape Town. This line consists of a fleet of 14 trains and 18 stations. In Case 1, 14 scenarios were simulated with each representing an incremental addition of a new train. Scenario 0 was the base case and simulated the current fleet of 14 old trains. Scenario 14 simulated the ideal future fleet which consists of 14 new trains.

An additional set of scenarios for Case 2 were created to test the effect on punctuality when increasing the general reliability of the whole system by 50% and incrementally adding new trains to the fleet.

There are therefore two different ways in which reliability was improved:

1. The incremental addition of new rolling stock
2. The improvement of the whole system's reliability

The effect on the performance parameters with the addition of new rolling stock can be found by analysing the two cases individually, however the effect of the improvement in the whole system's reliability can only be analysed by comparing Case 1 and Case 2.

A general note to consider is that the Number of delays and Average delay duration parameters were not proven valid and a -19.6% and +28.5% maximum error respectively, can be expected in the results presented here.

7.1 Case 1

The total number of delays decreased by a total of 29% from Scenario 0 to 14. From data it was found that rolling stock related delays contributed to 26% of all the delays logged. The 29% total reduction in delays is therefore a predictable result for the same reason that the new trains are assumed to be 100% reliable.

The total minutes delay decreased by a total of 37% from Scenario 0 to 14. This large decrease is a result of the compounding effect of 11% reduction in delay duration and 29% reduction in number of delays.

Mean delay duration decreased by a total of 11% from Scenario 0 to 14. This was because Rolling Stock was found to be the department with largest average delay duration, and since the new trains

were assumed to be 100% reliable (i.e. no rolling stock related delays), it can be expected that mean delay durations will decrease with the addition of new trains.

7.2 Case 2

Case 2 shows similar results with, total number of delays decreasing by 31% (2% more than in Case 1), total minutes delay decreasing by 36% (1% less than in Case 1) and mean delay duration decreasing by 7% (4% less than in Case 1).

7.3 Case 1 and Case 2 comparison

When each corresponding scenario for Case 1 and Case 2 is compared with each other the following differences were found:

- 13% reduction in number of delays from Case 1 to Case 2
- 19% reduction in total minutes delay from Case 1 to Case 2
- 6% reduction in mean delay duration from Case 1 to Case 2

Considering that in Case 2 the whole system was assumed to be 50% more reliable, it would seem that these differences are rather small. With the addition of the new trains the performance measures decreased as expected, however the same could not be said of when the system's reliability was improved. It was expected that if the probability that a delay-causing failure could occur was reduced by 50%, the performance measures of Case 2 would be 33% less than in Case 1. To understand why this is not the case, it must be understood as to how the two ways of improving reliability was implemented.

The addition of new rolling stock was simply replacing old trains with a high probability of experiencing a rolling stock related delay with a new train that is 100% reliable. Once the whole fleet was replaced the probability of a rolling stock related delay was 0% and therefore produced a predictable result.

The system reliability improvement was enforced by lengthening the frequency of primary delays by 50%. This only meant that the fleet of trains will experience 33% less primary delays, but not necessarily 33% less primary and secondary delays. Additionally, delays are measured only at the destination station, and therefore regardless of if a train experiences 2 or 3 delays during a trip, it will be counted as one delay and the duration will be the sum of the delays. The results are therefore a diluted representation of the system's reliability improvement.

The last finding indicates to what extent the punctuality of the service could be improved if the whole fleet is replaced and the rest of the system's reliability is improved by 50% through capitalisation and refurbishment. This was found by comparing the performance parameters of Case 1, Scenario 0 to Case 2, Scenario 14. The results reflecting the possible improvement that can be made for one direction of the line, over 6 months and are as follow:

- Total number of delays is reduced by 39% (1 453 delays)

- Total delay minutes is reduced by 47% (29 934 min)
- Mean delay duration is reduced by 13% (0:02:15)

These are significant improvements that can be made. The challenge however is to utilise the capital allocated by the government for this modernisation program effectively and efficiently. This can only be done if proper project planning is in place to prioritise capital spending and to ensure operational readiness before new technologies are introduced. This model assumed the new trains to be 100% reliable while operating on an aged and unreliable infrastructure. This is however unsustainable since the new trains are designed for a functional and well maintained infrastructure. It is therefore PRASA's mandate to provide not only an improved service by use of new trains, but also to ensure that the improved service is sustained long into the future.

In conclusion, to assist PRASA with this mandate, this model provides insight into the dynamic relationship between railway system improvement and passenger service improvement in terms of punctuality. It can therefore be used to support decision making with regards to capital expenditure for service improvement.

8. Recommendations

The following recommendations are made with regard to further research and introduction of the new rolling stock fleet.

8.1 Further research

It is recommended that this *Anylogic* model be expanded to include the down line, as well as the Kapteinsklip and Sarepta lines. The model will however have to be improved in terms of the validity of the Number of delays and Average delay duration, and computational time. A function must also be created that will allow the model to run pre-specified scenarios continually without user intervention between scenarios. This will allow more scenarios to be tested in a shorter time.

It is also recommended that a full agent based model be built of this exact line. This will allow the train to have acceleration and deceleration abilities. It will also localise the train's reliability. A delay will therefore not be enforced by signals but rather by the train's inherent capability to fail and stop at any point on its route. The model will then be much easier to calibrate since the train's behaviour is much closer to reality.

It will be very interesting to compare results of such a model with the one developed in this study. It will provide useful insight into new ways to simulate train fleet punctuality on a large scale.

8.2 Introduction of new trains

Introducing new trains one by one (as is currently the plan) may have socio-economic advantages (e.g. improved commuter comfort), however it is not recommended. If the new trains are to run at their design speeds, it will imply that a new schedule be created every time a new train is introduced. It will also have to be considered that the new train will be in much higher demand from passengers which holds the risk of overcrowding. The other risk is that if the infrastructure is not operationally ready, the new trains will not provide an improved service. This can lead to vandalism of the new trains.

It is therefore recommended that the introduction of the new trains be postponed until the necessary maintenance and refurbishments are done to the current infrastructure. It is also better to replace a fleet on a specific route completely rather than to replace them incrementally. This will create a homogeneous fleet which is simpler to schedule and will immediately provide an improved service. The old trains that are replaced can then be reallocated to other lines to help meet demand.

9. References

- [1] **PRASA**, “Annual Report 2012 /13,” 2012.
- [2] **PRASA**, “PRASA modernisation program,” 2015.
- [3] **RSR**, “State of Safety Report,” 2014.
- [4] **D. Peters**, Speech by the Minister of Transport, Ms Dipuo Peters, MP, on the annual general meeting of the Passenger Rail Agency of South Africa (PRASA), Hatfield, Pretoria, 2014.
- [5] **I. Solomons**, “PRASA train project on track,” *Engineering News*, Nov-2014.
- [6] **PRASA Rail operations**, Modernisation program presentation to Transport and Infrastructure portfolio: No public transport without PRASA, 2015.
- [7] **Alstom**, “X ’ TRAPOLIS TM Mega Product Sheet.” Alstom, Saint-Ouen, Cedex France, pp. 5–6, 2016.
- [8] **E. Kozan and A. Higgins**, Modeling Train Delays in Urban Networks, *Transp. Sci.*, vol. 32, no. 4, pp. 346–357, 1998.
- [9] **D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris**, *Fundamentals of Queueing Theory*. 2008.
- [10] **T. Huisman, R. J. Boucherie, and N. M. Van Dijk**, A solvable queueing network model for railway networks and its validation and applications for the Netherlands, *Eur. J. Oper. Res.*, vol. 142, no. 1, pp. 30–51, 2002.
- [11] **J. Yuan and I. a. Hansen**, Optimizing capacity utilization of stations by estimating knock-on train delays, *Transp. Res. Part B Methodol.*, vol. 41, no. 2, pp. 202–217, 2007.
- [12] **L. E. Meester and S. Muns**, Stochastic delay propagation in railway networks and phase-type distributions, *Transp. Res. Part B Methodol.*, vol. 41, no. 2, pp. 218–230, 2007.
- [13] **A. de Kort, B. Heidergott, R. J. van Egmond, and G. Hoogheimstra**, Train Movement Analysis at Railway Stations Procedures & Evaluation of Wakob’s Approach, 1st ed. Delft: Delft University Press, 1999.
- [14] **R. L. Burdett and E. Kozan**, A sequencing approach for creating new train timetables, vol. 32, no. 1. 2010.
- [15] **A. D’Ariano, D. Pacciarelli, and M. Pranzo**, A branch and bound algorithm for scheduling trains in a railway network, *Eur. J. Oper. Res.*, vol. 183, no. 2, pp. 643–657, 2007.
- [16] **F. Corman, A. D’Ariano, D. Pacciarelli, and M. Pranzo**, A tabu search algorithm for rerouting trains during rail operations, *Transp. Res. Part B Methodol.*, vol. 44, no. 1, pp. 175–192, 2010.
- [17] **A. D’Ariano, F. Corman, D. Pacciarelli, and M. Pranzo**, Reordering and Local Rerouting Strategies to Manage Train Traffic in Real Time, *Transp. Sci.*, vol. 42, no. 4, pp. 405–419, 2008.
- [18] **A. Higgins, E. Kozan, and L. Ferreira**, Heuristic techniques for single line train scheduling, *J. Heuristics*, vol. 3, no. 1, pp. 43–62, 1997.
- [19] **D. Goldberg and J. Holland**, Genetic Algorithms and Machine Learning, *Mach. Learn.*, vol. 3, pp. 95–99, 1988.
- [20] **J. W. Chung, S. M. Oh, and I. C. Choi**, A hybrid genetic algorithm for train sequencing in the Korean railway, *Omega*, vol. 37, no. 3, pp. 555–565, 2009.
- [21] **M. F. Gorman**, An application of genetic and tabu searches to the freight railroad operating plan problem, *Ann. Oper. Res.*, vol. 78, pp. 51–69, 1998.
- [22] **A. Nash and D. Huerlimann**, Railroad simulation using OpenTrack, *Comput. Railw. IX*, pp. 45–54, 2004.
- [23] **A. Radtke and D. Hauptmann**, Automated planning of timetables in large railway networks using a microscopic data basis and railway simulation techniques, *Adv. Transp.*, vol. 15, pp. 615–625, 2004.
- [24] **T. Schlechte, R. Borndörfer, B. Erol, T. Graffagnino, and E. Swarat**, Micro-macro transformation of railway networks, *J. Rail Transp. Plan. Manag.*, vol. 1, no. 1, pp. 38–48, 2011.
- [25] **C. C. Hwang and J. R. Liu**, A Simulation Model for Estimating Knock-on Delay of Taiwan Regional Railway, vol. 8, no. 1999, 2010.
- [26] **D. Middelkoop and M. Bouwman**, Simone: Large scale train network simulations, in *2001 Winter Simulation Conference*, 2001, no. 2, pp. 1605–1612.

- [27] **Anylogic**, “Anylogic_CSX case studies,” *Anylogic Case Studies*, 2014. [Online]. Available: <http://www.anylogic.com/case-studies/csx-solves-railroad-operation-challenges-with-and-without-anylogic-rail-library>. [Accessed: 15-Jan-2017].
- [28] **T. K. Ho, B. H. Mao, Z. Z. Yuan, H. D. Liu, and Y. F. Fung**, Computer simulation and modeling in railway applications, *Comput. Phys. Commun.*, vol. 143, no. 1, pp. 1–10, 2002.
- [29] **N. M. van Dijk**, Hybrid combination of queueing and simulation, in *2000 Winter Simulation Conference*, 2000, pp. 147–150.
- [30] **A. Azadeh, S. F. Ghaderi, and H. Izadbakhsh**, Integration of DEA and AHP with computer simulation for railway system improvement and optimization, *Appl. Math. Comput.*, vol. 195, no. 2, pp. 775–785, 2008.
- [31] **M. H. Dingler, Y.-C. Lai, and C. P. L. Barkan**, Impact of Train Type Heterogeneity on Single-Track Railway Capacity, *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2117, no. 1, pp. 41–49, 2009.
- [32] **M.-C. Shih, C. T. Dick, S. L. Sogin, and C. P. L. Barkan**, Comparison of Capacity Expansion Strategies for Single-Track Railway Lines with Sparse Sidings, *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2448, pp. 53–61, 2014.
- [33] **S. Sogin, C. P. L. Barkan, and M. R. Saat**, Simulating the effects of higher speed passenger trains in single track freight networks, *Proc. - Winter Simul. Conf.*, pp. 3679–3687, 2011.
- [34] **S. L. Sogin, C. T. Dick, Y.-C. Lai, and C. P. L. Barkan**, Analyzing the Incremental Transition from Single to Double Track Railway Lines, *Proc. Int. Assoc. Railw. Oper. Res. 5th Int. Semin. Railw. Oper. Model. Anal.*, pp. 1–20, 2013.
- [35] **Anylogic**, “Anylogic_Aurizon case study,” *Anylogic Case Studies*, 2012. [Online]. Available: <http://www.anylogic.com/case-studies/rail-yard-capacity-modeling>. [Accessed: 15-Jan-2017].
- [36] **L. Lategan**, Rekenaar-gesteunde ontwikkeling van ’n reëlmaatrooster vir voorstedelike treindienste, 2013.
- [37] **M. Gylee**, Punctuality Analysis-A Basis for Monitoring and Investment in a Liberalized Railway System, in *Proceedings of seminar held at the 22nd PTRC European Transport Forum*, 1994, p. 153–165.
- [38] **P. Rietveld, F. R. Bruinsma, and D. J. van Vuuren**, Coping with unreliability in public transportation chains: a case study for Netherlands, *Transp. Res. Part A* 35, pp. 539–559, 2001.
- [39] **P. D. F. Conradie**, Quantifying System Reliability in Rail Transportation in an Aging Fleet Environment, 2015.
- [40] **N. O. E. Olsson and H. Haugland**, Influencing factors on train punctuality — results from some Norwegian studies, vol. 11, pp. 387–397, 2004.
- [41] **A. Rudnicki**, Measures of regularity and punctuality in public transport operation, *Transp. Syst. Prepr. Eighth Int. Fed. Autom. Control*, 1997.
- [42] **NetworkRail**, Performance, *Public Performance Measure*, 2015. [Online]. Available: <http://www.networkrail.co.uk/about/performance/>. [Accessed: 09-Nov-2015].
- [43] **R. M. P. Goverde**, Punctuality of Railway Operations and Timetable Stability Analysis, 2005 .
- [44] **A. M. Law and W. D. Kelton**, *Simulation Modeling and Analysis*, Third. McGraw-Hill, 2000.

Appendix A1 – Timetables and trip times

Week day departure schedule		
Chris Hani	Kayelitsha	Phillipi
04:20:00		
		05:01:00
05:05:00		
05:20:00		
05:35:00	05:35:00	
05:50:00	05:50:00	
06:05:00		
	06:10:00	
06:15:00		
		06:16:30
06:30:00		
		06:41:00
06:42:00		
		06:45:00
	06:52:00	
06:55:00		
	07:10:00	
07:15:00		
		07:16:00
07:25:00		
07:35:00		
		07:45:00
07:50:00		
08:00:00		
08:20:00		
08:40:00		
09:10:00		
09:40:00		
10:10:00		
10:45:00		
11:15:00		
11:45:00		
12:15:00		
12:45:00		
13:20:00		
13:50:00		
14:25:00		
14:55:00		
15:25:00		
15:50:00		
16:15:00		
16:33:00		
17:08:00		
17:40:00		
18:05:00		
18:25:00		
19:00:00		

Saturday departure schedule		
Chris Hani	Kayelitsha	Phillipi
04:20:00		
05:05:00		
05:50:00		
07:25:00		
07:50:00		
08:20:00		
09:10:00		
09:40:00		
10:10:00		
10:45:00		
14:10:00		
14:40:00		
15:10:00		
15:50:00		
16:30:00		
17:10:00		
17:50:00		
18:10:00		
18:40:00		
19:22:00		

Sunday departure schedule		
Chris Hani	Kayelitsha	Phillipi
05:20:00		
05:50:00		
06:00:00		
06:05:00		
06:40:00		
07:20:00		
08:00:00		
08:40:00		
09:40:00		
10:40:00		
11:40:00		
12:40:00		
14:40:00		
15:40:00		
16:40:00		
17:30:00		
18:20:00		
19:10:00		

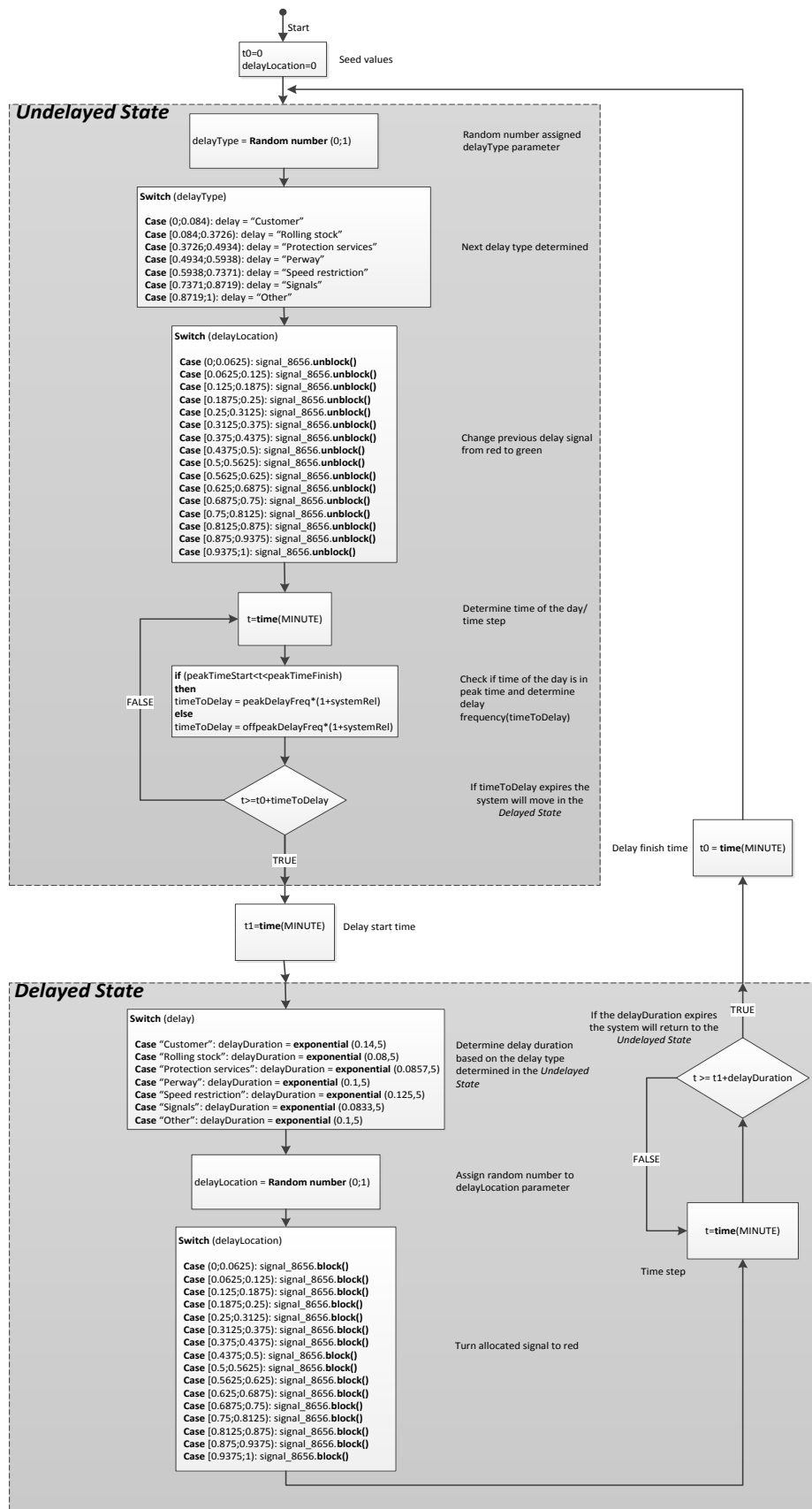
Train	Start Station	Time	End station	Time	Trip time
1	Chris Hani	04:20:00	Cape Town	05:23:00 AM	01:03:00
2	Kapteinssklip	05:01:00	Cape Town	05:39:30 AM	00:38:30
3	Chris Hani	05:05:00	Cape Town	06:07:30 AM	01:02:30
4	Chris Hani	05:20:00	Cape Town	06:21:00 AM	01:01:00
5	Khayelitsha	05:35:00	Cape Town	06:28:30 AM	00:53:30
6	Chris Hani	05:35:00	Cape Town	06:37:30 AM	01:02:30
7	Khayelitsha	05:50:00	Cape Town	06:43:30 AM	00:53:30
8	Chris Hani	05:50:00	Cape Town	06:48:00 AM	00:58:00
9	Chris Hani	06:05:00	Cape Town	06:56:00 AM	00:51:00
10	Khayelitsha	06:10:00	Cape Town	07:03:30 AM	00:53:30
11	Chris Hani	06:15:00	Cape Town	07:08:00 AM	00:53:00
12	Kapteinssklip	06:16:00	Cape Town	07:14:00 AM	00:58:00
13	Chris Hani	06:30:00	Cape Town	07:19:00 AM	00:49:00
14	Kapteinssklip	06:41:00	Cape Town	07:23:30 AM	00:42:30
15	Chris Hani	06:42:00	Cape Town	07:31:00 AM	00:49:00
16	Kapteinssklip	06:45:00	Cape Town	07:40:00 AM	00:55:00
17	Khayelitsha	06:52:00	Cape Town	07:45:00 AM	00:53:00
18	Chris Hani	06:55:00	Cape Town	07:54:00 AM	00:59:00
19	Khayelitsha	07:10:00	Cape Town	07:58:00 AM	00:48:00
20	Chris Hani	07:15:00	Cape Town	08:03:30 AM	00:48:30
21	Kapteinssklip	07:16:00	Cape Town	08:15:00 AM	00:59:00
22	Chris Hani	07:25:00	Cape Town	08:23:30 AM	00:58:30
23	Chris Hani	07:35:00	Cape Town	08:25:00 AM	00:50:00
24	Kapteinssklip	07:45:00	Cape Town	08:38:00 AM	00:53:00
25	Chris Hani	07:50:00	Cape Town	08:53:00 AM	01:03:00
26	Chris Hani	08:00:00	Cape Town	09:03:00 AM	01:03:00
27	Chris Hani	08:20:00	Cape Town	09:23:00 AM	01:03:00
28	Chris Hani	08:40:00	Cape Town	09:43:00 AM	01:03:00
29	Chris Hani	09:10:00	Cape Town	10:13:00 AM	01:03:00
30	Chris Hani	09:40:00	Cape Town	10:43:00 AM	01:03:00
31	Chris Hani	10:10:00	Cape Town	11:13:00 AM	01:03:00
32	Chris Hani	10:45:00	Cape Town	11:48:00 AM	01:03:00
33	Chris Hani	11:15:00	Cape Town	12:18:00 PM	01:03:00
34	Chris Hani	11:45:00	Cape Town	12:48:00 PM	01:03:00
35	Chris Hani	12:15:00	Cape Town	01:18:00 PM	01:03:00
36	Chris Hani	12:45:00	Cape Town	01:48:00 PM	01:03:00
37	Chris Hani	13:20:00	Cape Town	02:23:00 PM	01:03:00
38	Chris Hani	13:50:00	Cape Town	02:53:00 PM	01:03:00
39	Chris Hani	14:25:00	Cape Town	03:28:00 PM	01:03:00
40	Chris Hani	14:55:00	Cape Town	03:58:00 PM	01:03:00
41	Chris Hani	15:25:00	Cape Town	04:28:00 PM	01:03:00
42	Chris Hani	15:50:00	Cape Town	04:53:00 PM	01:03:00
43	Chris Hani	16:15:00	Cape Town	05:18:00 PM	01:03:00
44	Chris Hani	16:33:00	Cape Town	05:36:00 PM	01:03:00
45	Chris Hani	17:08:00	Cape Town	06:11:00 PM	01:03:00
46	Chris Hani	17:40:00	Cape Town	06:43:00 PM	01:03:00
47	Chris Hani	18:05:00	Cape Town	07:08:00 PM	01:03:00
48	Chris Hani	18:25:00	Cape Town	07:28:00 PM	01:03:00
49	Chris Hani	19:00:00	Cape Town	08:03:00 PM	01:03:00

Appendix A2 – Delay causes

Department	Delay cause
Customer Services	Passenger Related - Overcrowding of trains
Electrical	Bonding - Bonding failure
	Capital Works - Projects
	Endemic Faults - Any
	Maintenance - Emergency
	Maintenance - On track maintenance machines
	Occupations - Exceeded Time
	Signals - Signal power failure/supply
	Traction power - Panto Hookups
	Traction power - Traction power failure/supply
Facilities	Capital Works - Projects
	Endemic Faults - Any
Perway	Block joints - Faulty Block joints
	Foreign Objects - Block Joints
	Instructions - Temporary Speed Restrictions
	Maintenance - Emergency
	Maintenance - Material Defects
	Occupations - Exceeded Time
	Occupations - Planned Occupation
	Points - Dry Chair
	Rails - Alignment faults
	Rails - Broken Rails
Protection services	Assault/Robbery - Assault on Train
	Assault/Robbery - Robbery on Train
	Cable Theft - Signal
	Cable Theft - Traction wire
	Collisions / Derailments - Level crossing
	Collisions / Derailments - Person Struck
	Escorts - Waiting Escort
	Investigations - Protection Service Investigations
	Passenger Related - Drug Related Incidents
	Passenger Related - Injured / Sick passengers
	Passenger Related - Other reasons
	Passenger Related - Passengers outside trains
	Passenger Related - Passengers traveling between coaches
	Sabotage / Theft / Vandalism - Electrical
	Sabotage / Theft / Vandalism - Other reasons
	Sabotage / Theft / Vandalism - Perway
	Sabotage / Theft / Vandalism - Rolling Stock
	Sabotage / Theft / Vandalism - Signals

Department	Delay cause
Public	Civil Commotion - Demonstrations
Rolling stock	Endemic Faults - Any
	Mandatory Modifications - Intersite/SARCC Contract Action
	Points - Faulty Points
	Signals - Signal Failures
	Track circuits / Axles Counters - Faulty
	Track circuits / Axles Counters - Track Detection
	Train Sets - Faulty doors (sliding & cab)
	Train Sets - Late Ex SFF
	Train Sets - Motor Coach/Loco Defects
	Train Sets - Panto Hookups
	Train Sets - Set Compilation
	Train Sets - Unavailability
Signals	Cable Faults - Cable faults
	Capital Works - Projects
	Foreign Objects - Objects in Points
	Points - Faulty Points
	Signals - Faulty Signal Equipment
	Signals - Signal Failures
	Track circuits / Axles Counters - Faulty
	Track circuits / Axles Counters - Track Detection
Operations	Driver Problems - Availability of Personnel
	Driver Problems - Combi
	Driver Problems - Operating Irregularities
	Driver Problems - Roster Compiler
	Driver Problems - SPAD
	Driver Problems - Time lost by driver
	Guard Problems - Availability of Personnel
	Guard Problems - Operating Irregularities
	Guard Problems - Roster Compiler
	Marshaling Yard Delays - Operating Irregularities
	Marshaling Yard Delays - Section on Train
	Operating Office Irregularities - Wrong / None Reporting
	Personnel Issues - Availability of Personnel
	Planning Office Irregularities - Change / Poor schedule
	TCO Problems - Availability of Personnel
	TCO Problems - Operating Irregularities
	TCO Problems - Poor operating arrangements
Other	Conditions - Moisture/Skidding
	Launch & New Services - Any
	Weather - Frost
	Weather - Strong winds

Appendix B – Model algorithms



Switch (delayLocation)

```

Case (0;0.0625): signal_1366.unblock() && signal_1366_new.unblock()
Case [0.0625;0.125): signal_1656.unblock() && signal_1656_new.unblock()
Case [0.125;0.1875): signal_2066.unblock() && signal_2066_new.unblock()
Case [0.1875;0.25): signal_3656.unblock() && signal_3656_new.unblock()
Case [0.25;0.3125): signal_4156.unblock() && signal_4156_new.unblock()
Case [0.3125;0.375): signal_4556.unblock() && signal_4556_new.unblock()
Case [0.375;0.4375): signal_6036.unblock() && signal_6036_new.unblock()
Case [0.4375;0.5): signal_6066.unblock() && signal_6066_new.unblock()
Case [0.5;0.5625): signal_6356.unblock() && signal_6356_new.unblock()
Case [0.5625;0.625): signal_6666.unblock() && signal_6666_new.unblock()
Case [0.625;0.6875): signal_6856.unblock() && signal_6856_new.unblock()
Case [0.6875;0.75): signal_7256.unblock() && signal_7256_new.unblock()
Case [0.75;0.8125): signal_756.unblock() && signal_756_new.unblock()
Case [0.8125;0.875): signal_8056.unblock() && signal_8056_new.unblock()
Case [0.875;0.9375): signal_8356.unblock() && signal_8356_new.unblock()
Case [0.9375;1): signal_8656.unblock() && signal_8656_new.unblock()

```

Switch (delayLocation)

```

Case (0;0.0625): signal_1366.block() && signal_1366_new.block()
    if (delay == "Rolling stock")
    then signal_1366_new.unblock()
Case [0.0625;0.125): signal_1656.block() && signal_1656_new.block()
    if (delay == "Rolling stock")
    then signal_1656_new.unblock()
Case [0.125;0.1875): signal_2066.block() && signal_2066_new.block()
    if (delay == "Rolling stock")
    then signal_2066_new.unblock()
Case [0.1875;0.25): signal_3656.block() && signal_3656_new.block()
    if (delay == "Rolling stock")
    then signal_3656_new.unblock()
Case [0.25;0.3125): signal_4156.block() && signal_4156_new.block()
    if (delay == "Rolling stock")
    then signal_4156_new.unblock()
Case [0.3125;0.375): signal_4556.block() && signal_4556_new.block()
    if (delay == "Rolling stock")
    then signal_4556_new.unblock()
Case [0.375;0.4375): signal_6036.block() && signal_6036_new.block()
    if (delay == "Rolling stock")
    then signal_6036_new.unblock()
Case [0.4375;0.5): signal_6066.block() && signal_6066_new.block()
    if (delay == "Rolling stock")
    then signal_6066_new.unblock()
Case [0.5;0.5625): signal_6356.block() && signal_6356_new.block()
    if (delay == "Rolling stock")
    then signal_6356_new.unblock()
Case [0.5625;0.625): signal_6666.block() && signal_6666_new.block()
    if (delay == "Rolling stock")
    then signal_6666_new.unblock()
Case [0.625;0.6875): signal_6856.block() && signal_6856_new.block()
    if (delay == "Rolling stock")
    then signal_6856_new.unblock()
Case [0.6875;0.75): signal_7256.block() && signal_7256_new.block()
    if (delay == "Rolling stock")
    then signal_7256_new.unblock()
Case [0.75;0.8125): signal_756.block() && signal_756_new.block()
    if (delay == "Rolling stock")
    then signal_756_new.unblock()
Case [0.8125;0.875): signal_8056.block() && signal_8056_new.block()
    if (delay == "Rolling stock")
    then signal_8056_new.unblock()
Case [0.875;0.9375): signal_8356.block() && signal_8356_new.block()
    if (delay == "Rolling stock")
    then signal_8356_new.unblock()
Case [0.9375;1): signal_8656.block() && signal_8656_new.block()
    if (delay == "Rolling stock")
    then signal_8656_new.unblock()

```


Appendix C – Published article

MATHEMATICAL AND SIMULATION TECHNIQUES FOR MODELLING URBAN TRAIN NETWORKS

N. Wilson¹, C.J. Fourie^{1*} & R. Delmistro²

ARTICLE INFO

Article details

Submitted by authors 16 Sep 2015
Accepted for publication 24 May 2016
Available online TBC

Contact details

* Corresponding author
cjf@sun.ac.za

Author affiliations

- 1 Department of Industrial Engineering
Stellenbosch University, South Africa
- 2 Group Strategy Office
PRASA, South Africa

DOI

<http://dx.doi.org/10.7166/XX-X-1364>

ABSTRACT

Railway systems can pose complex problems for the scheduling and operation of trains. A passenger rail service's first priority is to provide a punctual and safe transport service to its customers. But doing so is a major challenge for rail network operators, as disruptions are inevitable, especially in densely-populated networks. Disruptions can be caused not only by infrastructure or rolling stock breakdowns, but also by maintenance activities, new rolling stock, or new train services. Managing these disruptions and predicting the extent of its effects is a crucial part of rail network operation. Mathematical models and simulation can be applied to these problems. This paper will review the literature concerning the modelling of train networks.

OPSOMMING

Spoorweg stelsels skep soms komplekse probleme met betrekking tot skedulering en die bedryf van treine. 'n Passasiers-spoordiens se eerste prioriteit is om stiptelike en veilige vervoer te verskaf aan sy gebruikers. Om 'n stiptelike en betroubare diens te lewer is uiter aard 'n groot uitdaging vir netwerk operateurs, aangesien treindienste maklik ontwig word in dig bevolkte netwerke. Ontwigtinge word nie net deur infrastruktuur en rollende materiaal falings veroorsaak nie, maar ook deur infrastruktuur onderhoud, nuwe rollende materiaal, en nuwe treindienste wat ingestel kan word. Die bestuur van dié ontwigtinge en die akkurate vooruitskatting van die effek op die res van die netwerk is 'n kritiese komponent van die bedryf van 'n trein netwerk. Wiskundige modelle en simulasië metodes kan toegepas word op dié tipe probleme. Hierdie artikel sal dus die literatuur bespreek wat handel oor die modellering van trein netwerke.

1 INTRODUCTION

Railway network companies often need to model and simulate the operation of their trains. This need usually arises with the expansion or maintenance of infrastructure, or the addition of new rolling stock and services. Infrastructure expansion entails adding new links, stations, or additional lines on a specific route. Furthermore, perway, electrical, and signals maintenance all contributes to train operations being disrupted to some extent. And adding train services or new rolling stock requires major operations planning and rescheduling. Forecasting the effect on the operation of the network before the implementation of such changes is a crucial component of planning. Bottlenecks, line capacities, demand satisfaction, and delay propagations are all areas that need to be identified and calculated before large capital amounts are spent. This can be done through the use of mathematical models and simulation. The optimisation of existing operations can also be done using these tools.

2 OBJECTIVE

The objective of this paper is to review the literature on the different modelling techniques that are used to describe the operation of train networks. This will lay the groundwork for developing the most appropriate application of these techniques on whichever case study of train networks needs to be modelled by future research work. The two spectrums of modelling train networks, analytical models and simulation models, will be discussed. In section 2, mathematical models and heuristic algorithms will be discussed, while in section 3 simulation models will be covered.

2.1 Mathematical models and heuristics algorithms

Analytical models tend to be limited in scope and complexity, but they mostly form the basis on which simulation models are built. With the advances made in computer capabilities in the last 10 years, the use of analytical models has become scarce. Kozan and Higgins [1] developed an analytical model to estimate delays for individual trains and track links in an Australian rail network. They compared the results with those obtained from a simulation algorithm. For 93 per cent of the 157 scheduled trains, the analytical model's delay estimates were within 20 per cent of those of the simulation algorithm's estimates. This shows that if the scope of the model is small enough, analytical and simulation models can produce similar answers.

When it comes to optimising train schedules, heuristic algorithms are used, such as job shop, genetic, and Tabu-search algorithms. These heuristic algorithms will be discussed in later sections.

2.2 Queuing models

Queuing theory, which was originally referred to as 'telegraphic theory', was developed in the 1920s for telecommunication services. The application of this theory has since expanded to the computer, manufacturing, retail, services, and transport industries.

Queuing processes are usually described by six characteristics; these are listed by Gross *et al.* [2] as:

1. Arrival pattern of customers.
2. Service pattern of servers.
3. Number of service stages.
4. Number of service channels.
5. Queuing discipline.
6. Capacity of the system.

The arrival pattern in most queuing models is stochastic in nature, and follows a certain probability distribution of inter-arrival times. It can, however, also be deterministic, depending on the systems being modelled. When setting up the parameters for arrival, it is necessary to know if agents can arrive in bulk - i.e., simultaneously - and if so, the probability distribution of the size of the bulk. In some models, an agent can decide not to join the queue upon arrival; this is referred to as 'balked'. In some cases, an agent can enter a queue, and then lose patience after a while, and leave the queue; this is referred to as 'reneged'. Another case may be when there is more than one queue and an agent switches from one queue to another; this is called 'jockeying'. Further on, when an arrival distribution does not change over time, it is referred to as 'stationary'; and when it does change, it is called 'nonstationary'. Note that jockeyed and reneged arrivals are not considered in rail systems. Trains cannot arrive in bulk because of headway constraints forcing trains to have a certain time or distance buffer between them. Similarly, trains cannot renege or jockey in a queue (waiting track) if the driver becomes impatient. It is possible, however, for a train to balk. When a serious disruption occurs on a route, oncoming trains can be rerouted where possible, or even be cancelled.

Similar to arrival patterns, service patterns also have distributions describing the time an agent spends being serviced. Agents can also be serviced in bulk or individually. The service time can, however, be influenced by the size of the queue or arrival pattern. In such a case, it is referred to as a 'state-dependent service', but generally arrival and service patterns are assumed to be independent [1]. Another aspect of service time, as with arrival patterns, is that it may change over time - e.g., when learning takes place and the service process becomes quicker and more efficient. The same terms mentioned previously, 'stationary' and 'nonstationary', are used for

such service processes. This is not usually applicable in rail systems, as trains have specified dwell times at stations.

How an agent is chosen for service from a queue is referred to as the queuing discipline. The most common discipline is the first-come-first-served (FCFS) principle; however, in some inventory systems, the last-in-first-out (LCFS) principle applies. Other systems have priority schemes that are usually called either 'pre-emptive' or 'non-pre-emptive'. Pre-emptive priority is when a high priority agent enters a queue, the service on a low priority agent is paused, and the high priority agent is serviced first. In the case of a non-pre-emptive priority, the high priority agent will be moved to the front of the queue, but will only be serviced when the agent being served at that moment is finished. Passenger rail systems mostly work on the FCFS principle, whereas freight rail systems might have different disciplines that take into account the importance of the freight content.

Some systems have limited queues, which creates a limited system capacity, such as a doctor's waiting room with a limited number of chairs. On the other hand, some queuing systems have infinite capacity, as in the case of judicial processes or waiting lists. In the case of rail systems where stations and sections are the servers, queues are limited.

Queuing systems can have more than one service channel. In general, it is preferred to have a single queue feeding multiple channels - e.g., customs at airports and railway stations with more than one platform. This usually applies in systems where the agents have no preference about which service channel they want to use. On the other hand, in systems like most supermarkets with multiple tills, customers line up in multiple queues.

The last aspect of queuing systems is stages of service. Systems may have more than one service stage; manufacturing systems are good examples of this. Parts will, for instance, be assembled and then moved forward to be checked for quality. If the quality is not satisfactory, the assembly will be fed back to the previous stage, or else the assembly will move forward to be painted. Passenger rail systems only have one service stage, while freight trains may have more (e.g., freight being unloaded and then the train moving to the hump yard).

The following points summarise queuing systems:

1. An agent arrives according to a certain probability distribution or fixed inter-arrival time.
2. The agent then enters or does not enter the queue, depending on the type of system.
3. The agent then moves from the queue to get serviced for a duration specified by the modeller. This can be for a stochastic or fixed time period.
4. After the agent is serviced, it leaves the system and the next agent in the queue is serviced, depending on the queuing discipline.

Huisman *et al.* [3] developed a queuing network model to compute the long-term performance of rail networks. To achieve this, a decomposition of the network and its detailed components was necessary. These components include stations, junctions, and sections. The network performance was measured by the mean delay and delay probability of the trains arriving at their destinations. Because train movements are not known over the long term, assumptions were made to simplify the modelling of stations. One of the assumptions is to model the storing tracks outside of the model. Thus, when a train finishes its route, it exits the model and is stored in a queue outside the model. The halting track is where the train starts its route and where the passengers alight or board the train. The next train can only enter the model after the train on the halting track has departed.

The occupation times at the halting tracks are assumed to be distributed exponentially and to be equal for all train types. The stations are modelled as multi-server queuing systems (since stations have more than one platform), with Poisson arrival distributions.

The same principles were applied to junctions and sections, except that these were single server queues. If a junction is occupied, the next train falls into the queue, until the junction is clear. This occupation time is also distributed exponentially.

Sections were broken up into signal blocks, with each block acting as a separate queuing system. Bottlenecks and delays were then calculated by adding up all the waiting times in the queues. These waiting times were compared with the practical delay times of the trains.

Table 1: Queueing notation [1]

Queueing notation A/B/X/Y/Z			
Characteristic		Symbol	Explanation
A B	Inter-arrival time distribution	M	Exponential
		D	Deterministic
	Service time distribution	E_k	Erlang type ($k=1, 2, 3, \dots$)
		H_k	variety of k-exponentials
		PH	Phase type
		G	General
X	Number of parallel servers	1, 2, ∞	
Y	System capacity	1, 2, ∞	
Z	Queue discipline	FCFS	First come, first served
		LCFS	Last come, last served
		RSS	Random selection
		PR	Priority
		GD	General

The model showed good accuracy, even though the timetable was not taken into account. Yuan and Hansen [4] and Meester and Muns [5] have both emphasised the lack of queuing models to consider timetables, since they are reliant on probability distributions for inter-arrival times. Moreover, fixed arrival and departure times were also not considered, and the impact of speed variations was neglected. Huisman *et al.* [3] instead suggested a way to capture speed variances among different train types by ignoring block (signalling) sections in a section between stations. However, the model does include one block section before and after each station, to ensure that trains do not arrive in bulk at stations. This means that, for instance, if a section has five signalling blocks, the middle three sections will be removed from the model and only the first and last sections will be included. This allows enough distance for a train with a different speed to have a significant variance in free running time; here, free running time refers to the time a train takes to travel between stations without any disruptions. The model of Huisman *et al.* [3] was applied to two major lines of the Dutch network, Rotterdam to Utrecht, and Den Haag to Utrecht. The traffic on this network is extremely heterogeneous, with three different train types (implying three different train speeds) running three different services.

De Kort *et al.* [6] also applied a similar queuing model, based on Wakob's approach, to Den Hague station in the Netherlands. Wakob's approach breaks up all the components of a station and analyses them independently as separate queues. Arrival and service times are both assumed to fit an Erlang distribution, resulting in $E_k(\lambda)/E_t(\mu)/1$ queues for the whole queuing system. De Kort *et al.* [6] argued that service time variations should be dependent on running time and dwell time variations, instead of on independent probability distributions. It was found that this approach overestimates delays and, alternatively, models the 'worst case scenario'. This may be related to the fact that Wakob's approach returns the upper bound of the delay duration instead of the mean and standard deviation. Although this approach is inappropriate for delay propagation analysis, it can be useful for capacity planning purposes [6].

Queuing models can serve as a good alternative to simulation in order to estimate delays, although - as mentioned previously - modelling large networks becomes difficult to solve analytically. Kozan and Higgens [1] explain this complexity of train networks:

“A train network is complex in that it includes many intersections, uni- and bidirectional track links of various lengths, sidings, and track capacity. Train services vary with different upper velocities, slack time, scheduled stops, non-uniform departure times, and include train connections as described in the introduction of the paper. In the case of train connections and intersections, a train can suffer a delay from another that is scheduled much earlier and from a different part of the network.”

“As well, the distribution of arrival times for each train at any station or intersection depends on the distribution of current delay, which can be different for each train service. Hence, delay to both the trains and at stations (or intersections) are interdependent. Therefore, the calculation of expected delay requires a solution of equations.”

2.3 Job shop models

Branch and bound algorithms have been used to develop and optimise timetables. These models transform train networks into large job shop models. Typically, trains will be jobs and stations and sections will be machines. In job shop models, a number of different jobs need to be completed by a number of machines. A job will have a specified time and order that it has to spend at each machine. For example, Job A will use Machine 1 for two minutes, then Machine 2 for five minutes, and lastly Machine 3 for three minutes. Job B will first use Machine 2 for three minutes, then Machine 1 for five minutes, and end off with Machine 3 for one minute. Figure 1 shows an illustration of this simple model. It is important to note that each machine can only work on one job at a time. This means that when Job B is finished with Machine 2, Job A can move to Machine 2. Similarly, when Job A is finished with Machine 1, Job B can move to Machine 1. Whichever job finishes using Machines 1 and 2 first then moves to Machine 3. The other job will then have to wait for the first job to finish before moving to Machine 3. In the example illustrated in Figure 1, both jobs will arrive at Machine 3 at the same time. In such cases, priority rules can be implemented. Problems of this nature create the need to determine what the optimal sequence of machine use is; i.e., which job should use which machine when. Branch and bound algorithms are used to solve these problems. For further explanations of job shop models and branch and bound algorithms, refer to Gross *et al.* [2].

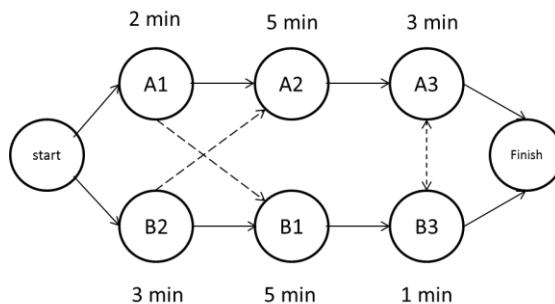


Figure 1: Simple job shop model

Rail networks can be similarly modelled, where trains are seen as jobs, and stations, sections, and junctions are seen as machines. There are, however, key differences between train network models and classical job shop models. These are listed as follows [7]:

- Jobs and machines do not have lengths as do trains and sections.
- While moving from one section to another, a train's 'head' will occupy the next section, while the 'tail' will occupy the current section. A train may thus occupy two sections at a time, whereas jobs can normally only occupy one machine.
- Train acceleration, deceleration, and cruising speed for a specific section cannot always be pre-defined, since it is dependent on the train in front.
- Trains can visit sections more than once, whereas jobs are mostly assumed to visit machines only once.

- Passing facilities such as passing loops on rail sections are equivalent to capacitated buffers or parallel machines. These are very difficult features to model with a standard job shop model.

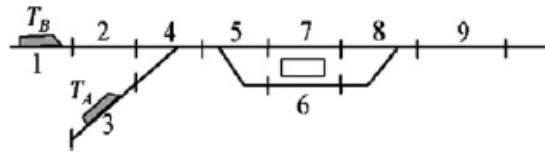


Figure 2: Small network with block sections [8]

In Burdett and Kozan's [7] paper, the authors explain how these differences were incorporated in order to produce realistic results. D'Ariano *et al.* [8] developed a job shop model for the Dutch railway network. Figure 2 shows a small network on which the model in Figure 3 is based. Note that each block section is represented by a machine or a resource, as referred to in this paper, and Trains A and B are the jobs. A minimum headway of one signal block between trains is modelled and indicated by the dotted arrows in Figure 3. For example, Train A can only enter Block 5 when Train B has exited Block 7.

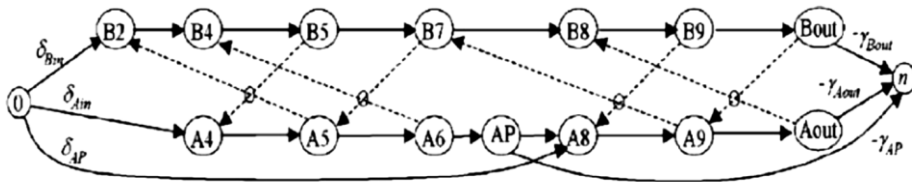


Figure 3: Job shop graph of two trains [8]

This model was expanded to model the Schiphol rail network, which includes the stations of Nieuw-Vennep, Hoofddorp, Amsterdam Lelylaan, and Amsterdam Zuid. The network consists of 86 block sections, 16 platforms, traffic in two directions, and 54 trains.

The model wished to solve the train scheduling problem for real-time rail network management. The objective function was to minimise the maximum secondary delays at all stations by all trains. It was found that these algorithms perform better than the dispatching rules commonly used in relation to average and maximum delays.

Burdett and Kozan [9] used a hybrid job shop model with time window constraints to solve the train scheduling problem when adding additional train services. In their later work [7], they again used the job shop approach, but then further refined the solution using simulated annealing and local search meta-heuristics. This allowed them to shift trains more easily and feasibly within the solution.

2.4 Tabu search

Tabu search is a meta-heuristic algorithm that memorises the most recent local optimum. As soon as a solution is found that is better than the previous best solution, the algorithm will store it and discard the previous best solution (i.e., the solution becomes tabu). This also implies that the algorithm will never return to the same solution twice. The tabu search thus eliminates the possibility for the search to get stuck on a local maximum, and continually searches for new local optima in the solution space.

Corman *et al.* [10] compare a tabu search algorithm with a local search algorithm and various hybrid algorithms previously developed [8,11] to solve routing and scheduling problems in the Dutch rail network. The study focused on a bottleneck at the dispatching area of Utrecht Den Bosch, which consists of 191 block sections, 21 platforms, and 50 kms of track. The algorithms had to search through 356 possible routes for the best solution. The results showed that the tabu search algorithm reached better solutions faster than did the other heuristic algorithms.

Similar conclusions about the quality and speed of solutions reached by tabu search methods were reached by Higgins *et al.* [12], who solved the problem of a single track line with occasional sidings for opposing trains to pass each other.

2.5 Genetic algorithms

Genetic algorithms are very effective and robust algorithms to determine global optima. Gradient-based methods, such as Steepest Ascent, Conjugate Gradient, or Lagrangian Multiplier, usually converge faster to local optima or a local optimum than a genetic algorithm. In cases of multimodal functions, however, they may miss the global optimum more often than not. Genetic algorithms are based on the theory of genetic evolution, where the fittest genes in a chromosome survive and the weakest genes die away in the process of reproduction. To put it differently, the offspring of two parent chromosomes will only consist of the best genes found in both parents. In this way, continual improvement in fitness takes place with every generation.

Considering the algorithm, each solution is represented by a chromosome. Stochastic mutation of some of these offspring is brought in at pre-determined instances in order to make sure the algorithm does not get stuck on a local optimum. The numerical values of a solution's parameters are converted to a series of binary digits, and each parameter is then represented by a gene. When a gene thus evolves, the digits of its binary code change to either 1 or 0 [13].

Genetic algorithms are not commonly used for solving train scheduling problems. However, Higgins *et al.* [12] used a genetic algorithm to solve a single line train scheduling problem. In their study, each gene contained three attributes: the delayed train, the train with the highest priority or right of way, and the track section where the conflict will occur. With each parent in this instance consisting of six genes (e.g., six train schedule solution), the fittest two parents are chosen to mate and produce two children with genes from both parents with a single randomly-selected crossover point. The genes before the crossover point are transferred to the first child, while the genes after the crossover point are transferred to the second child. Mutation in this algorithm has a very low probability, however, when mutation happens and the conflict gene changes, and the neighbouring genes also change. Changing only one conflict gene by mutation is not good in train scheduling problems [12]. The genetic algorithm in this study proved to outperform the tabu search and local search heuristics, which the authors also used to solve the same problem.

It seems that most of the cases where genetic algorithms were used were in cases of single track lines with traffic in both directions [3,14,15].

3 SIMULATION MODELS

Saayman and Bekker [16] explain simulation as an attempt to solve real world problems by first building a model that represents the current state and operation of a system as realistically as possible. This is achieved by making argued simplifications and assumptions. The model can then be used to solve, experiment with, or optimise the modelled system. Saayman and Bekker [16] explain further that simulation allows the modeller to include the stochastic nature of a real world system. It allows for big scopes and high complexity systems. It is difficult, however, to validate a model, since the whole point of simulation is to forecast the effects of change to a system before spending capital to implement the intended change. Model validation is usually done by comparing the 'current state' model with actual system behaviour. In this way, the modeller can make the assumption that the model is a realistic representation of the system. Simulation is thus a tool that should be applied with care, since getting answers is easy, but getting realistic answers is a fine skill [16].

Hwang and Liu [17] developed a simulation model to forecast the effect of increasing demand for railway capacity of the regional railway system in Taiwan. The idea was not only to model the increase in the line capacities, but to also improve the efficiency of the current capacity. The model's objective was the accurate estimation of knock-on delays (secondary delays) as a result of a primary delay. The following input parameters were used to represent the network:

- Railway condition: the line, stations, and track layouts of the stations.
- Traffic condition: minimum dwell time and scheduled timetable.

- Control condition: minimum headways, section capacity, and recovery time.

With these parameters, the model was run assuming no delays; i.e., strictly following the scheduled timetable. To determine the effect of a primary delay on the network, a delay event had to be created. This event or primary delay is defined by four parameters: location of delay, delay start time, delay release time, and the magnitude of the delay. The magnitude of the delay is simply the difference between the delay start time and the delay release time. The resulting secondary delays were thus one of the outputs of the model. These delays were then used to create a simulated timetable.

To validate their model, Hwang and Liu [17] used actual train operating data. The arrival-departure time data of a specific day was retrieved from the Centralised Train Control database of the Taiwan Railways Administration. Later, actual delay data was also collected in order to compare it with the simulation output. A route conflict delay was chosen as the real event that serves as the primary delay. The model proved to be within 120 seconds of the actual delay time 77.5 per cent of the time, and 62.5 per cent of the time it was within 60 seconds. Figure 4 shows the Marvey diagram of the normal timetable without any delays, and Figure 5 shows the diagram for the simulated timetable. It is clear that a delay occurred between Shongshan and Taipei stations, and that the next seven trains were affected by it.

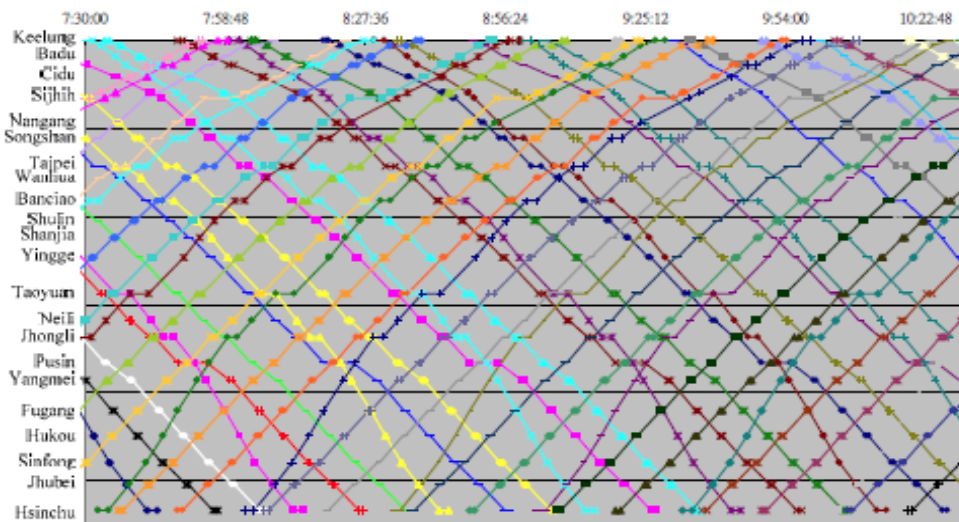


Figure 4: Normal timetable without delays [17]

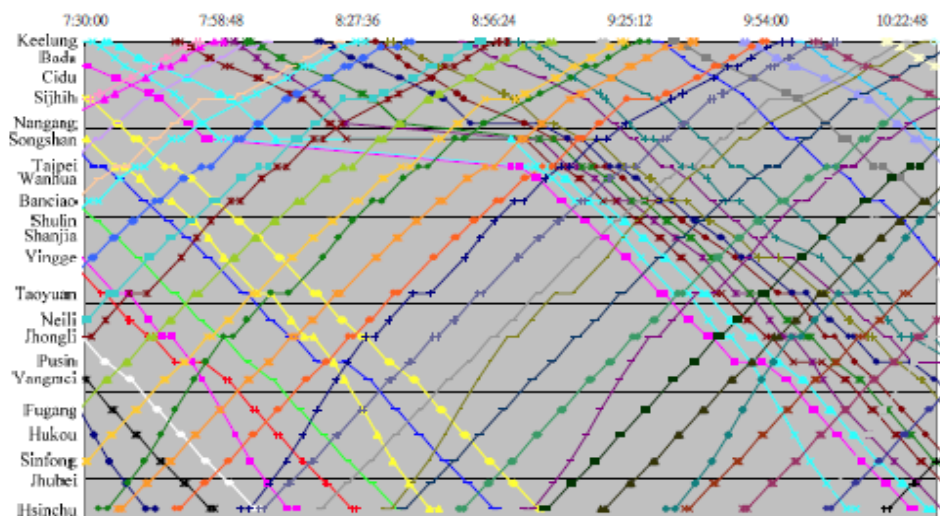


Figure 5: Simulated timetable diagram with delays [17]

Hwang and Liu [17] went further and compared different delay reduction strategies and how they influence the total secondary delays; the effect of three strategies are shown in Figure 6. It is interesting to note the exponential relationship between primary (or first delay) and secondary delays (or knock-on delays). This can be explained by the fact that the larger the primary delay is,

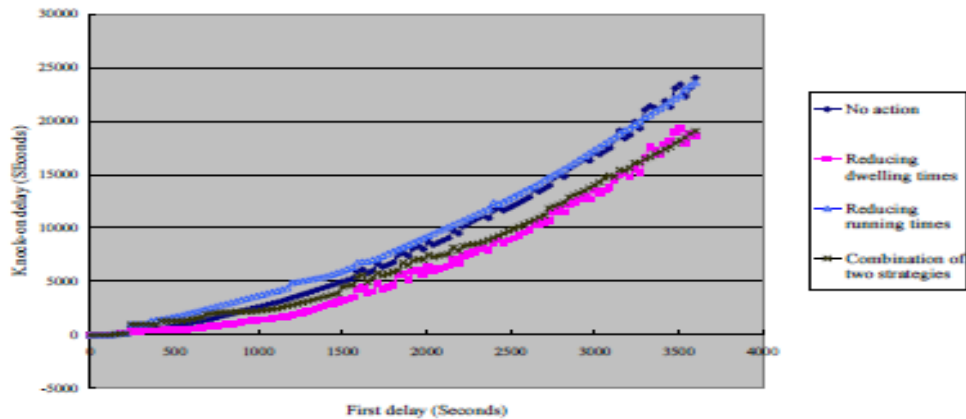


Figure 6: Total knock-on delays at the destination station [17]

the harder it is for a train to recover any of the lost time. A train is naturally limited by its ability to use these three strategies to recover the lost time created by the primary delay. A train has a minimum allowed dwell time at stations, and is also subject to speed limits on sections. These limitations thus translate into knock-on effects on later trains, which results in an exponential growth in the total delays.

Middelkoop and Bouwman [18] demonstrated the use of *Simone* simulation software to model the entire Dutch rail network. The software requires the following as inputs to the model:

- Infrastructure data.
- Timetable.
- Simulation-specific parameters.
- Network properties in relation to disruptions and disturbances.
- Operational rules.
- Statistical indicators for the simulation output.

The software then produces the indicators pre-specified by the user and an animation of the network operation. Figure 7 shows an example of the animation output that *Simone* produces. The figure shows a part of the Dutch rail network and all the trains operating on it, with the red circle indicating a highly congested part of the network. Each type of train has a unique colour. Most parts of the model were constructed by the software's automatic model generator. The model included 600 stations, 1,100 track sections, and 350 trains, which is significantly large. The model was able to show, for example, the punctuality of trains in certain parts of the network and the relationship between initial delays and the sum of delays (as done by Hwang and Liu [17]).

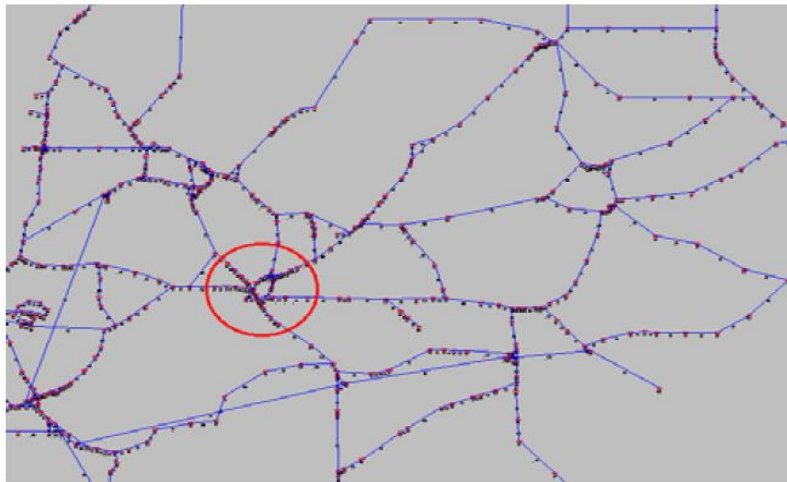


Figure 7: Simone simulation animation output [18]

Van Dijk [19] suggested that queuing theory and simulation can be combined. He argued that the advantages of queuing theory (e.g., generic components and few detailed data needed) reduce the disadvantages of simulation (i.e., high levels of complexity and the need for detailed data). In the same way, a simulation's advantages (i.e., real-life complexity and real-life uncertainties) reduce the queuing theory's disadvantages (i.e., over-simplification and unrealistic constraints).

Azadeh *et al.* [20] used a Visual SLAM coding language to develop a simulation model of a complex rail system consisting of 50 stations and both passenger and freight trains. An analytical hierarchy process (AHP) method was used to weight the qualitative and quantitative inputs and outputs, which were then converted to a data envelopment analysis (DEA). The objective of the model was to find ways to increase passenger train reliability and decrease the turn-around time of both passenger and freight trains.

Ho *et al.* [21] developed a general-purpose multi-train simulator that enables users to model without carrying out program code modifications. The simulator has been used in Hong Kong and China for studies of traffic control at conflict areas, scheduling optimisation, and the energy management of trains.

Train networks can be simulated in two ways. One is time-based modelling, where a time span is broken up into equal intervals and train movement is calculated at each interval. Although this is a very realistic representation of train movement, it requires a large amount of information with every update, which makes it computationally intensive. Time-based models are typically used in signalling layout design and energy consumption analysis [21].

The second way of simulating train movement is event-based. This method is similar to the queuing models discussed in Section 2.2. The train's movement is described in terms of a chain of events. For example, the train arrives at a station at a specified arrival rate and stays for a certain time period. The train then leaves and enters a track section, which marks the start of the next event. Each event's duration is characterised by a certain probability distribution. Although event-based models may reduce computational time significantly compared with time-based models, train movement updates are not synchronised between events [21].

4 CONCLUSION

This paper discussed the various ways to model and schedule train networks. First, purely analytical models were covered that showed that networks can be modelled accurately without advanced computational methods. They are, however, very limited in terms of scope and network complexity.

Second, heuristic methods were discussed. It can be concluded that these methods are very effective in optimising large complex networks. They allow the modeller to find global optima

amid a solution plane consisting of many local optima. Optimising train schedules for dense rail networks seem to be possible with the right combination of these heuristic algorithms.

Last, the use of simulation was discussed. Simulation allows for very large scopes and even entire networks to be modelled [18]. It also has the ability to include important infrastructure detail and simulate reality fairly accurately. Moreover, it possesses the ability to animate the model, making the complex nature of a rail network visual and easier to understand.

The challenge is to combine these mathematical modelling techniques and simulation software to represent and predict real-life situations as accurately as possible. For future work, it is suggested that these techniques be applied to a case of the Passenger Rail Agency of South Africa (PRASA). In this case, PRASA has to introduce new and faster trains into a homogeneous rail system. The rail traffic will then become heterogeneous, implying that the network will have to be re-scheduled. The other issue is the following question: On which routes and in what quantity should the new trains be introduced so that service reliability will improve? The answers to this question can be estimated with the use of simulation modelling. Since most advanced simulation software available uses discrete events to model systems, and train operations can easily be described by discrete events, it is proposed to use discrete event simulation. Once a validated model is developed, heuristic methods can then be used to optimise the operation of trains in very specific scenarios. A very clear objective function and constraints are necessary, however, which could lead to a reduction of scope.

REFERENCES

- [1] Kozan, E. and Higgens, A. 1998. Modeling train delays in urban networks. *Transp. Sci.*, 32(4), pp. 346-357.
- [2] Gross, D., Shortle, J.F., Thompson, J.M. and Harris, C.M. 2008. *Fundamentals of Queueing Theory*.
- [3] Huisman, T., Boucherie, R.J. and Van Dijk, N.M. 2002. A solvable queueing network model for railway networks and its validation and applications for the Netherlands. *Eur. J. Oper. Res.*, 142(1), pp. 30-51.
- [4] Yuan, J. and Hansen, I.A. 2007. Optimizing capacity utilization of stations by estimating knock-on train delays. *Transp. Res. Part B Methodol.*, 41(2), pp. 202-217.
- [5] Meester, L.E. and Muns, S. 2007. Stochastic delay propagation in railway networks and phase-type distributions. *Transp. Res. Part B Methodol.*, 41(2), pp. 218-230.
- [6] De Kort, A., Heidergott, B., Van Egmond, R.J. and Hoogheijstra, G. 1999. *Train movement analysis at railway stations: Procedures & evaluation of Wakob's Approach*. 1st Edition, Delft: Delft University Press.
- [7] Burdett, R.L. and Kozan, E. 2010. A sequencing approach for creating new train timetables, 32(1)..
- [8] D'Ariano, A., Pacciarelli, D. and Pranzo, M. 2007. A branch and bound algorithm for scheduling trains in a railway network. *Eur. J. Oper. Res.*, 183(2), pp. 643-657.
- [9] Burdett, R.L. and Kozan, E. 2009. Techniques for inserting additional trains into existing timetables. *Transp. Res. Part B Methodol.*, 43(8-9), pp. 821-836.
- [10] Corman, F., D'Ariano, A., Pacciarelli, D. and Pranzo, M. 2010. A Tabu search algorithm for rerouting trains during rail operations. *Transp. Res. Part B Methodol.*, 44(1), pp. 175-192.
- [11] D'Ariano, A., Corman, F., Pacciarelli, D. and Pranzo, M. 2008. Reordering and local rerouting strategies to manage train traffic in real time. *Transp. Sci.*, 42(4), pp. 405-419.
- [12] Higgins, A., Kozan, E. and Ferreira, L. 1997. Heuristic techniques for single line train scheduling. *J. Heuristics*, 3(1), pp. 43-62.
- [13] Goldberg, D. and Holland, J. 1998. Genetic algorithms and machine learning, *Mach. Learn.*, 3, pp. 95-99.
- [14] Chung, J.W., Oh, S.M. and Choi, I.C. 2009. A hybrid genetic algorithm for train sequencing in the Korean railway. *Omega*, 37(3), pp. 555-565.
- [15] Gorman, M.F. 1998. An application of genetic and tabu searches to the freight railroad operating plan problem. *Ann. Oper. Res.*, 78, pp. 51-69.
- [16] Saayman, S. and Bekker, J. 1999. Drawing conclusions from deterministic logistic simulation models. *Logist. Inf. Manag.*, 12(6), pp. 460-466.
- [17] Hwang, C.C. and Liu, J.-R. 2010. A simulation model for estimating knock-on delay of taiwan regional railway. 8(1999).
- [18] Middelkoop, D. and Bouwman, M. 2001. Simone: Large scale train network simulations. *2001 Winter Simulation Conference*, 2001(2), pp. 1605-1612.
- [19] Van Dijk, N.M. 2000. Hybrid combination of queueing and simulation. *2000 Winter Simulation Conference*, 2000, pp. 147-150.
- [20] Azadeh, A., Ghaderi, S.F. and Izadbakhsh, H. 2008. Integration of DEA and AHP with computer simulation for railway system improvement and optimization. *Appl. Math. Comput.*, 195(2), pp. 775-785.
- [21] Ho, T.K., Mao, B.H., Yuan, Z.Z., Liu, H.D. and Fung, Y.F. 2002. Computer simulation and modeling in railway applications. *Comput. Phys. Commun.*, 143(1), pp. 1-10.

