

Cost functions for railway operations and their application to timetable optimisation

Thesis submitted to University College London for the degree of
Doctor of Philosophy

by

Aris Pavlides

Department of Civil, Environmental & Geomatic Engineering

Centre for Transport Studies

University College London

October 2016

Declaration

I, Aris Pavlides, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

Abstract

This thesis investigates cost functions for evaluating and optimising the performance of a timetable with mixed train services. Specifically, the performance considered herein includes crowdedness, journey time, punctuality and waiting time. In order to examine the implications of optimising using these cost functions, a multi-objective optimisation algorithm is developed to derive an optimised timetable for mixed train services. The optimisation algorithm consists of three stages: a Genetic Algorithm (GA) is used to determine the optimal sequence of train runs, followed by Dijkstras shortest path algorithm for determining the optimal schedule based on the sequence determined by GA, and finally an iterative Hill-Climbing procedure for determining the optimal number of train runs in the system. Experiments were carried out on the Brighton Main Line and examined the effect of different timetabling parameters. The first series of experiments showed that the cost of the timetable can be driven down simply through resequencing the trains such that trains exiting the network quickly are more evenly distributed through the time period examined. This occurs due to the fact that trains exiting early create a buffer which can absorb delays, preventing their propagation. The experiments have also shown that different demand levels influence the number of trains to be scheduled. The optimal number of trains to schedule though relies on the equilibrium between the crowdedness and punctuality cost function. Scheduling additional trains leads to a non-linear reduction in the marginal gains in terms of the crowdedness function while, on the other hand, the cost of punctuality increase exponentially. Finally, we derive the Pareto Frontiers for different combinations of cost functions. This research contributes to the state-of-art of railway system analysis and optimisation.

Acknowledgements

First of all I would like to thank my principal supervisor Dr Andy Chow for his patience and incredible support he has provided me with for the duration of my PhD. The discussions we had not only provided me with important feedback on my work but have also helped me become a better researcher by helping me realise the importance of being inquisitive not only in the academic world but in life in general. I would also like to thank my industrial supervisor, Ms Catherine Baker, for all the time she devoted in helping me in my project by not only providing me with constructive feedback but also by helping me understand the British railway industry in order to better understand the implications of my work in a real life context.

I would also like to thank other people who have contributed in successfully completing my project. Special thanks must be given to Creon for the endless hours he spend helping me code my model. His input in helping me progress my coding skills has been truly invaluable. Students E-Yang Tan, Konrad Bablinski, Qi Zhang and Tao Tao have all provided me with work that has, in some form or another, been implemented in my project. The members of the Railway Research Group are too many to be named individually but have all helped me by providing me with feedback on the progress on my work while also helping me to better understand how the railway industry operates.

This project would have never been possible in the first place without the tremendous support of my family. My parents, Pavlos and Yianna as well as my siblings Kostis and Antonia have always been next to me, always ready to unselfishly support me in every way imaginable. I will forever be indebted to them.

I would like to convey my warm and profound thanks to my dearest friends in the United Kingdom for their patience and invaluable support throughout my study period. I am grateful to my friends and fellow colleagues Sascha, Stelios, Rui,

Alexandra, Arash and Savvas. Michalis Konstantoulakis deserves my gratitude for being the single most important reason for giving me the motivational boost needed to pursue my dream of obtaining a PhD.

The list would not be complete without thanking Florentia for being by my side, supporting and motivating me to finish my thesis.

Participating in UCL's Integrated Engineering program has been a tremendously enjoyable and rewarding experience which was enhanced through my collaboration with amazing people. Dr Abel Nyamapfene was the person who gave me the chance to participate in the integrated Engineering program and was always there looking after me. Furthermore, Dr Stuart Grey and Dr Paul Groves were a joy to work with, creating a friendly atmosphere in which I could carry out my duties.

Finally, I would like to thank the people in Network Rail and RSSB for providing me with the necessary tools and data needed for my case studies for the Brighton Main Line and East Coast Main Line.

Contents

Abstract	v
Acknowledgements	vii
List of Figures	xiii
List of Tables	xv
Abbreviations	xvii
Notation	xix
1 Introduction	1
1.1 Timetabling in the British railway industry	1
1.2 Project motivation	4
1.3 Research objectives	5
1.4 Thesis overview	7
2 Literature Review	9
2.1 Introduction	9
2.2 Cost functions for railway timetable optimisation	10
2.2.1 Cost functions for multi-objective timetable optimisation	10
2.2.2 Performance metrics for railway timetabling	16
2.3 Optimisation algorithms	29
2.4 Summary	31
3 Cost Functions	35
3.1 Introduction	35

3.2	Specification of timetable and associated constraints	36
3.3	Performance metrics and cost functions	40
3.3.1	Running times of trains	41
3.3.2	Waiting times of passengers	42
3.3.3	Punctuality of service	44
3.3.4	Crowdedness	47
3.4	The cost of travel time savings	47
3.4.1	Values of travelling and waiting	50
3.4.2	Punctuality multipliers	51
3.4.3	Crowdedness multipliers	52
3.4.4	Cost function formulations	55
3.5	Summary	55
4	Optimisation of a railway timetable	59
4.1	Introduction	59
4.2	Description of the optimisation algorithm	60
4.2.1	First stage - Genetic Algorithm	63
4.2.2	Second stage - Dijkstra's Algorithm	68
4.2.3	Third stage - Hill-Climbing Algorithm	71
4.3	Passenger calculations	73
4.3.1	Origin-destination matrix	73
4.3.2	Derivation of train demand	75
4.4	Punctuality modelling	78
4.5	Model validation	85
4.5.1	East Coast Main Line	85
4.5.2	Algorithm convergence - ECML network	89
4.5.3	BRaVE simulation environment	90
4.5.4	Validation of timetable construction method in BRaVE	92
4.5.5	Validation of the optimisation procedure	96
4.6	Summary	102
5	Case study	105
5.1	Introduction	105
5.2	Brighton Main Line	106

5.3	Algorithm convergence - BML network	110
5.4	Optimised sequence process	112
5.4.1	Optimised sequence traits - Presentation of results	113
5.4.2	Optimised sequencing traits - Discussion	115
5.5	Impact of scheduling additional trains	118
5.5.1	Impact of overcrowding - Presentation of results	119
5.5.2	Impact of overcrowding - Discussion	128
5.6	Pareto analysis	131
5.6.1	Pareto analysis - Presentation of results	136
5.6.2	Pareto analysis - Discussion	143
5.7	Summary	145
6	Conclusions	149
6.1	Thesis overview	149
6.2	Contribution to the research field	156
6.3	Future work	158
A	Origin-destination matrix for the ECML	163
B	Timetables generated for validation	165
C	Origin-destination matrix for the BML	167
D	Terminology	169
	References	175

List of Figures

2.1	Railway capacity parameters [87]	19
3.1	Timetabling variables illustration	37
3.2	Signalling block example occurrence	39
3.3	Punctuality cost function	44
4.1	Flowchart of the optimisation algorithm	62
4.2	Genetic algorithm flowchart	65
4.3	Sample network	70
4.4	Sample network for origin-destination matrix illustration	74
4.5	Construction of the stochastic matrices flowchart	80
4.6	Network for time matrix illustration	81
4.7	Time-space diagram for time matrix illustration	81
4.8	ECML between Alexandra Palace and Hatfield	87
4.9	Initial iterations of the optimisation algorithm	89
4.10	Algorithm convergence after 400 iterations	90
4.11	Validation process flowchart	92
4.12	Monitoring point recording the arrival of trains for the optimisation algorithm (A) and BRaVE (B)	95
4.13	Passenger matrix used for algorithm validation	97
5.1	Brighton Main Line between Gatwick Airport and Brighton	107
5.2	Origin-destination matrix for a 100% loading factor	109
5.3	Convergence with 400 iterations	111
5.4	Optimised train sequencing for low demand levels	113
5.5	Optimised train sequencing for average demand levels	114
5.6	Optimised train sequencing for high demand levels	115
5.7	Optimised train sequencing illustration	116
5.8	Optimised number of trains for low demand	120
5.9	Optimised number of trains for average demand	120
5.10	Optimised number of trains for high demand	121
5.11	Marginal timetable improvements for the low demand scenario	122
5.12	Marginal timetable improvements for the average demand scenario	123
5.13	Marginal timetable improvements for the high demand scenario	123
5.14	Optimised number of trains for low demand - Timetable break down	124

5.15 Optimised number of trains for average demand - Timetable break down	124
5.16 Optimised number of trains for high demand - Timetable break down	125
5.17 Changes in the cost of crowdedness when additional trains are scheduled	126
5.18 Changes in the cost of journey time when additional trains are scheduled	126
5.19 Changes in the cost of punctuality when additional trains are scheduled	127
5.20 Changes in the cost of waiting time when additional trains are scheduled	127
5.21 Pareto frontier for crowdedness against punctuality - Average demand	137
5.22 Pareto frontier for crowdedness against punctuality - High demand .	137
5.23 Pareto frontier for crowdedness against journey time - Average demand	139
5.24 Pareto frontier for crowdedness against journey time - High demand	139
5.25 Pareto frontier for punctuality against journey time - Low demand .	140
5.26 Pareto frontier for punctuality against journey time - Average demand	141
5.27 Pareto frontier for punctuality against journey time - High demand	141

List of Tables

1.1	Key measures concerned by different railway stakeholders [20]	5
3.1	Monetary coefficients [85]	51
3.2	Lateness multipliers [78]	52
3.3	Sitting penalties for crowdedness [86]	53
3.4	Standing penalties for crowdedness [86]	53
4.1	Parent chromosomes for crossover	64
4.2	Offsprings after crossover	64
4.3	Permutation sequencing	66
4.4	Origin-Destination matrix example	74
4.5	Passengers boarded example	75
4.6	Demand matrix example for four stations	76
4.7	Total seats offered example	77
4.8	Time matrix example	81
4.9	Train mix between 08:00-09:00 on a weekday	88
4.10	Timetable excerpt from the optimisation algorithm	93
4.11	Timetable excerpt from BRaVE	93
4.12	Possible train sequences for algorithm validation	98
4.13	Timetable cost for all sequences using brute force	98
4.14	Timetable cost for the number of trains scheduled (Brute force)	100
4.15	Timetable cost for the number of trains scheduled (Optimisation algorithm)	100
4.16	Timetable cost per sequence and optimal number of trains (Brute force)	101
4.17	Timetable cost per sequence and optimal number of trains (Optimisation Algorithm)	101
5.1	Train mix between 08:00-10:00 on a weekday	107
5.2	Passenger mix for the time period 08:00 - 10:00	108
5.3	Sequencing results for different demand levels	114
5.4	Optimised number of trains for low demand - Solution table	122
5.5	Optimised number of trains for average demand - Solution table	125
5.6	Optimised number of trains for high demand - Solution table	128
5.7	Pareto Frontier construction for Crowdedness and Punctuality	133

5.8	Pareto Frontier construction for Crowdedness and Journey Time . .	134
5.9	Pareto Frontier construction for Punctuality and Journey Time . .	135
5.10	Pareto Frontier to be constructed	135
A.1	Origin-destination matrix between Alexandra Palace and Hatfield .	164
B.1	Arrival times as generated by the optimisation procedure	166
B.2	Arrival times as generated by BRaVE	166
C.1	Origin-destination matrix between Gatwick Airport and Brighton .	168

Abbreviations

BML	B righton M ain L ine
BRaVE	University of B irmingham R ailway V irtual E nvironment
DEA	D ata E nvelopment A nalysis
DfT	D epartment for T ransport
ECML	E ast C oast M ain L ine
IM	I nfrastucture M anager
IVT	I n V ehicle T ime
OD matrix	O igin D estination matrix
KPI	K ey P erformance I ndicator
ORR	O ffice of R ail and R oad (formerly Office of Rail Regulation)
VoT	V alue of T ime (also known as the value of travel time savings)

Notation

$A_{i \rightarrow j}$	total number of seats offered by the train services from station i to station j
a_n	seating capacity of train n
c^y	monetary cost coefficient attributed to each cost function for passenger type y
$D_{n,i}$	dwell time of train n at station i
$D_{n,i}^*$	minimum dwell time of train n at station i given the regulations by Network Rail
h_b^*	minimum headway requirements (in minutes along) block section b
L_n	the length of train n
N	the trains to be scheduled ($n \in N$)
N_i	the set containing all the trains which traversed station i ($N_s \subseteq N$)
$p_{n,i}^y$	number of passengers of type y on board train n as it travels between station i and j
$R_{n,i}^y(p, a_n)$	crowdedness time multiplier for passenger type y in the train n travelling from station i to station j
S	the stations in the network ($((i, j) \in S)$)
S_n	the set of all stations visited by train n ($S_n \subseteq S$)
$T_{n,i \rightarrow j}$	running time of train n from station i to station j
$t_{n,b}^{in}$	the time at which block section b is marked as occupied by train n
$t_{n,b}^{out}$	the time at which block section b is marked as being released from train n
V_n^*	maximum speed for train n given the speed limit and train characteristics

$V_{n,b}$	speed of train n in block section b
$\Delta_{i,j}$	distance between station i and station j
$\delta_{n,b}$	sighting distance of the signal at the entrance of block b by the driver of train n
$\lambda_{i \rightarrow j}^y(t)$	arrival rate of passengers of type y going from station i to station j (function of time)
$\sigma_{n,i \rightarrow j}$	departure time of train n from station i going to station j
$\tau_{n,b}$	arrival time of train n at block section b
Φ	time deviation from the timetable which is allowed for the train to still be considered on-time

Chapter 1

Introduction

1.1 Timetabling in the British railway industry

The British railway industry, which is the oldest in the world, has experienced a vast increase in usage since its privatisation in 1994. At the same time nonetheless, there has been a dramatic surge in the complexity of its organisational structure, underscoring inter alia the importance of having efficient timetabling procedures in place [20, 53, 64]. Currently, such procedures are viewed to be overly lengthy and thus in merit of further streamlining to optimise the required involvement of the various stakeholders in the railway industry.

Every five years, the government defines the level of service expected from the railway industry and determines the level of public expenditure. The government

then enters into a series of negotiations with Network Rail (the infrastructure manager) and the Office of Rail and Road (Network Rail's economic regulator) to determine the requirements in terms of system capacity and its reliability [20, 44]¹. The final set of specifications are formalised in the High Level Output Statement (HLOS) which defines the performance targets for the railway sector during the five-year period [29].

At the moment, there are two ways for train operators² to gain access to the railway network: purchasing specific slots in the timetable (known as open access operations) and purchasing the right to run contracted services on given parts of the network (known as franchises). Most of the services operating on the network right now are franchises [12]. The process of franchising starts with the Government defining the performance targets a franchise should meet by referring to the specifications formalised in HLOS. Operators then submit bids which are evaluated by the Government during the passenger franchising process [12]. Once a bid is accepted, a franchise agreement is signed between the Government and the train operator which binds the operator to provide a railway service for the agreed period [65]. A list with all franchises along with their expiration dates can be found in [75]. Freight operators are not legally bound to provide a specific franchise and are not subject to the performance standards which apply for train operators. The freight market is only governed by freight customers but freight operators must

¹System capacity is measured as the number of passengers and freights the network can accommodate while reliability refers to the percentage of services arriving at their destination on time [29]

²The term 'train operators' refers to the companies which operate passenger trains. On the other hand, freight operators are only responsible for operating freight services

still liaise with Network Rail to gain access to the infrastructure [44].

The process of franchising and bidding described above leads to the production of the static timetable which can be broken down to two processes [20, 44]. The Long Term Planning (LTP) process produces two timetables per year; one in December and one in May [20, 44]. These timetables are being devised 28 weeks before their introduction and are being made available to the public 12 weeks before their introduction in order to give time to passengers to plan their journey in advance [20]. The Short Term Planning (STP) process has the purpose of scheduling trains which missed the LTP deadlines and also considers the impact of engineering works by Network Rail [20]. The STP planning process is initiated 18 weeks before the timetable's introduction but, like the LTP timetable, the STP timetable is being made available 12 weeks before its actual date of introduction [20]. Changes can still be made to the timetable even on the day before the actual implementation, but these changes usually concern freight trains rather than passenger trains [20].

The final timetable must comply with the rules set for each one of the ten available routes which exist in the UK [65]. The Timetable Planning Rules are route-specific guidelines which are devised by Network Rail and the operators and provides the set of rules that the timetable should abide by [44]. These rules provide information including minimum headway requirements, timing allowances, dwell times etc. [34]

1.2 Project motivation

The increased complexity of the British railway structure along with the significant increase in the traffic it attracts, have led to the creation of inefficiencies in the industry [53]. The 2013 Rail Technical Strategy [72] has identified four key areas for improvement for British railways: reduced *carbon*, increased *capacity*, decreased *operating costs* and improved *customer satisfaction*. These criteria have now become known as the 4C.

Timetabling construction has been identified as one of the areas upon which British railways can improve on in order to meet the targets set by the 4C. At the moment, railway timetabling in the UK is a manual process which aims to produce feasible timetables with no consideration being paid on whether the final timetable is optimal [20]. The Future Traffic Regulation Optimisation (FuTRO) project, funded by the Rail Safety and Standards Board (RSSB), aims to develop an optimisation framework which can be used to construct a railway timetable which will be optimal in terms of the 4C criteria. Following the publication of the objectives of FuTRO, Chen and Roberts [20] have stated the performance metrics for assessing a railway timetable as well as the stakeholders for which each performance metric is relevant (Table 1.1).

In its final form, the project can be used by the rail industry ³ to inform full development of optimisation algorithms for use within the timetable planning and

³In the context of this project, the term 'rail industry' refers to Network Rail, Train Operating Companies, Freight Operating Companies and the Rail Safety and Standards Board

TABLE 1.1: Key measures concerned by different railway stakeholders [20]

	Transport Volume	Travel Time	Connectivity	Punctuality	Resilience	Comfort	Energy	Resource Usage
Infrastructure Manager	*	*		*	*		*	*
Train Operators		*	*	*		*	*	*
Railway Customers		*	*	*		*		
Government	*						*	

traffic management systems, taking into account the impact on different stakeholders. This will ensure that the railway timetable produced will contribute in meeting the targets set by the 4C.

1.3 Research objectives

One of the objectives of the project is to identify the performance metrics applicable to railway timetables and provide their mathematical formulations. Rather than focusing on the performance metrics relevant to a single stakeholder in the railway industry, the interests of multiple stakeholders will be considered. This will fill a gap in literature which, up to now, only focuses on the simultaneous optimisation of two or three objective functions that are usually tailored according to the needs of just a single stakeholder (e.g. [11, 32, 88, 89]).

Following the identification of the performance metrics, the next step is to transform the metrics such that they have the same dimension; enabling for the estimation of a timetable's total 'cost'. This is a novelty since in the literature, when authors optimise under different objectives, either the objectives have the same

dimension (e.g. [11, 88, 89]) or multi-objective optimisation techniques are used (often ϵ constraints which constrain all objectives but one which is the one optimised) which avoid the problem of dealing with differently dimensioned objectives (e.g. [4, 37]). Even though such techniques may be effective when dealing with a couple of objectives, when the number of objectives increases, their effectiveness suffers. Consequently, since in this project more than two objectives will be used to evaluate a timetable, a different approach is required.

An important aspect of the project is to understand how sensitive the optimal solution is to a range of different parameters. Such parameters are:

- In what ways does the off-peak hours optimal solution differ from the peak hours solution.
- Is there a way to sequence the trains such that the new sequence leads to lower timetable cost.
- What impact (if any) does the passenger mix have on the optimal solution.
- How do different multi-objective optimisation techniques influence the optimal solution.

The above does not represent an exhaustive list of the parameters to be examined but rather provides the foundation upon which the experiments can be carried out. Such an analysis is important since, to the best of the author's knowledge, no such analysis has been carried before and will help to shed light into the many different factors which may influence the quality of railway timetables.

The main purpose of the project can therefore be summarised as the formulation of a set of cost functions which capture the performance of a railway timetable, and the analysis of how the optimal decision changes if any timetabling parameters vary.

1.4 Thesis overview

The thesis is structured as follows: Chapter 2 provides the literature upon which this research is based on to calculate the cost functions for British railway operations. Algorithms which are often used to solve the timetabling problem will also be given.

In Chapter 3, the formulation of the constraints used to construct a feasible timetable is given and their formulation explained. The objective functions to be used in the optimisation problem are also provided by first identifying performance metrics to evaluate a timetable's performance, and then a monetary cost is associated to these metrics to transform them to cost functions.

Chapter 4 explains the optimisation algorithm developed which will act as the main tool for carrying out the analysis. The optimisation model is then validated by using it in conjunction with a simulation environment and comparing the output.

Chapter 5 presents a case study based on the Brighton Main Line and more specifically the section between Gatwick Airport and Brighton. The case study examines

different timetabling parameters and how they impact on the optimal solution.

The Pareto Frontiers for different combinations of cost functions are also constructed.

Finally, Chapter 6 concludes the thesis and outlines the project's future steps.

Chapter 2

Literature Review

2.1 Introduction

A review of the current literature dealing with the performance metrics used for railway timetable optimisation is provided in this chapter. Literature on multi-objective railway timetabling is examined to identify the different optimisation techniques used by various authors as well as the objectives used to evaluate timetable performance.

This chapter is organised as follows: Section 2.2 examines the different performance metrics that have been employed over the years to assess the effectiveness of a railway timetable. A summary is also provided which examines the different performance metric combinations which have been employed by different authors

to formulate the multi-objective train timetabling problem. Finally, Section 2.3 provides an overview into the different optimisation models developed over the years to tackle the railway timetabling problem.

2.2 Cost functions for railway timetable optimisation

An extensive literature currently exists which aims to optimise a railway timetable given a set of cost functions. Section 2.2.1 provides the existing literature on the different techniques used by authors when optimising timetables under different objectives. Section 2.2.2 describes the various methodologies which have been developed over the years to assess timetable-related performance metrics. These metrics have been divided into four broad categories: network and system capacity, journey time, punctuality and waiting time

2.2.1 Cost functions for multi-objective timetable optimisation

As of the time of writing, a number of existing studies examine a railway timetable using more than one objective. Abril et al. [1] analyse the trade-off between network capacity and punctuality by adding buffer time in the timetable but, even

though the impact on network capacity is shown, the improvements in punctuality are not quantified. Yuan and Hansen [96] analyse the trade-off between network capacity and punctuality as well but, rather than timetabling, they calculate network capacity as a function of the train frequency in critical components within a given time interval. This train frequency is a function of running times, buffer time, supplements and dwell time [96]. Goverde et al. [40] examine how capacity utilisation and delays are impacted under different signalling systems. Gibson et al. [38] have used empirical data to establish a relationship between network utilisation and the delay of all trains over a section by fitting a non-linear regression of the form:

$$D_{it} = A_i \exp(\beta C_{it}) \quad (2.1)$$

The term D_{it} is defined as the total reactionary delay of all trains over a line i during time interval t . Gibson et al. [38] define reactionary delays as the extent to which an operator's trains delay another operator. A_i and β are the section specific and route specific constants respectively and C_{it} is the utilisation of section i during time t . The analysis of the empirical data suggests that as the section's utilisation increases, reactionary delays increase exponentially. This formulation is less likely to be of relevance to networks with minimal utilisation levels. This is because when a primary delay occurs in such networks, the large time period between services is very likely to absorb the delay, preventing it from delaying any subsequent trains. Hallowell and Harker [41] have run simulations where the delay in a timetable is examined for three different traffic levels (low, average and

high traffic). In general, they show that increases in traffic levels lead to higher delays but they also report several cases where the standard deviation of delays seems to be inversely proportional to the level of traffic. Even though Harker and Hallowell [41] attribute this fact to the difference in the number of simulated train movements between traffic volume levels, it may also be attributed to the fact that they only consider two stations in the experiments (origin and destination). Also the fact that they use different weights for the delays of each train type might have also had an impact on their experiments.

The term D_{it} is defined as the total reactionary delay of all trains over a line i during time interval t . Gibson et al. [38] define reactionary delays as the extent to which an operator's trains delay another operator. A_i and β are the section specific and route specific constants respectively and C_{it} is the utilisation of section i during time t . The analysis of the empirical data suggests that as the section's utilisation increases, reactionary delays increase exponentially. Hallowell and Harker [41] have run simulations where the delay in a timetable is examined for three different traffic levels (low, average and high traffic). In general, they show that increases in traffic levels lead to higher delays but they also report several cases where the standard deviation of delays seems to be inversely proportional to the level of traffic. Even though Harker and Hallowell [41] attribute this fact to the difference in the number of simulated train movements between traffic volume levels, it may also be attributed to the fact that they only consider two stations in the experiments (origin and destination). Also the fact that they use different weights for the delays of each train type might have also had an impact on their

experiments.

Peterson [74] evaluates the trade-off between journey time and punctuality by redistributing allowance time in a pre-constructed timetable.

Bussieck et al. [11] construct a timetable which minimises travelling time and waiting time for transfer passengers.

Albrecht [4] maximises average train loading and minimises the average time passengers spend on the platform waiting for their train to arrive but instead of constructing a timetable, the train frequency during a time interval is calculated.

Ghoseiri et al. [37] minimise fuel consumption and passenger travelling time.

Albrech et al. [2, 3] devise a control strategy for a single train to minimise energy consumption while ensuring that the journey time does not exceed a given limit.

In their work, Albrech et al. [2, 3] show that such a unique optimal control strategy exists and that it can be found within acceptable time frames.

Fuel consumption is also examined by Higgins et al. [42] who develop a timetable which also minimises delays. However, fuel efficiency maximisation during the timetabling process can be called into question due to the fact that accurate consumption rates require dynamic information (e.g. acceleration, deceleration) which is not available in static timetables.

Some authors consider more than two objectives in the timetabling process but

focus extensively only one aspect. For example, Dorfman and Medanic [32] and Li et al. [48] analyse a timetable using four criteria: total time to clear a line, total delays, maximum delay which can occur in the timetable and time efficiency of the timetable. This means that the above authors place a lot of emphasis on punctuality which is evaluated using different metrics.

A similar approach is adopted by Goverde et al. [40] and Sama et al. [80] who develop a multi-objective optimisation problem to minimise the impact of disturbance management through real-time rescheduling. The objectives considered are the following:

- Maximum tardiness - the maximum positive difference between a train's estimated and scheduled arrival time at any node in the network.
- Cumulative tardiness - calculated as the sum of all delays at all nodes in the network.
- Cumulative tardiness end - calculated as the sum of all delays at the time of their last operation ¹.
- Punctuality - the number of trains arriving late at their last operation ².
- Priority cumulative tardiness end - the sum of weighted delays associated with a train's last operation.
- Priority cumulative tardiness end cost - similar to above with the extensions of a delay threshold Φ and a penalty cost for each delay which occurs.

¹Last operation is defined by Sama et al. [80] as the time that a train enters or exits a network as well as the time a train stops at any intermediary nodes in its path

²Trains are considered late if they arrive at a node Φ minutes after its scheduled arrival time

- Scheduled deviation - penalises both early and late arrivals and penalises late departures at the nodes.
- Total completion - the sum of delays of all the trains at their last node in the network.
- Travel time - the sum of the time that all scheduled train spend in the network.

These objectives are then combined using an adaptation of Data Envelopment Analysis (DEA) which 'uses linear programming to determine the relative efficiencies of a set of homogeneous (comparable) units' [80]. For any feasible solution to the problem, DEA provides an efficiency score for each objective function, indicating how well each objective performs for the given solution [40]. The problem with DEA is that it provides an efficiency metric for each function individually rather than the optimisation problem as a whole so, in cases where an objective is preferred over the rest, this is not captured by the DEA.

Sameni and Preston [81] also use Data Envelopment Analysis to analyse the efficiency of railway operations since they recognize that some performance metrics have different units of measurement, making it difficult to compare with each other in a holistic way. The two performance metrics used by Sameni and Preston [81] are: the number of kilometres a timetable offers and delay minutes. The model though is likely to favour timetables with low frequency trains since trains will have more time between them, allowing the timetable to absorb delays but also

make sure that sufficient demand has been generated during that time, leading to high train loading values.

2.2.2 Performance metrics for railway timetabling

Throughout the literature, authors use different performance metrics to evaluate railway timetables. For example, the British railway industry has two ways of assessing delays: the number of services which arrive at their destination within three minutes and the second is the average delays of each train at each station in its path [60]. In literature, punctuality can be measured as the total delays by all trains or the maximum delay expected to occur [32]. This chapter has the purpose of presenting all the different methodologies used in the literature and industry to assess railway timetable performance metrics which cover the following areas: capacity, journey time, punctuality and waiting time.

Section 2.2.2.1 explains the complications behind the estimation of the capacity of a railway network as well as the different methods which attempt to calculate network capacity. It also provides a description of system capacity and relates it to train loading. Section 2.2.2.2 provides the literature for evaluating travelling time. The different methods for modelling train delays as well as the formulations used to calculate a timetable's performance in terms of the delays on arrival are given in Section 2.2.2.3. Finally, Section 2.2.2.4 analyses the literature for assessing a timetable's waiting time from the passengers point of view.

2.2.2.1 Capacity

In the railway industry, the term 'capacity' encompasses a wide range of definitions, creating the need to clarify what does the term 'capacity' really refer to. The British railway industry has two definitions for capacity. Network capacity is measured as the number of passenger and freight trains the network can accommodate while system capacity refers to the number of passengers or freights that a given timetable can serve [29].

Network capacity

Analysing railway capacity is important from the infrastructure manager's (IM) and train/freight operators' point of view since train/freight operators pay a fee to use the infrastructure to run their services [45]. Therefore, the maximum number of trains that can traverse through the network in a given period of time can serve as a benchmark to evaluate the performance of a railway timetable.

Static methods for calculating capacity, model the railway environment using mathematical formulae and calculate a value of capacity which represents the maximum number of trains that can traverse the network within the time period examined [10, 28, 56]. The downside of static approaches is that the decision variables refer to the maximum train services a railway network can support during time interval T without assigning a value to the entry/exit time of the trains from the locations they will visit. The absence of such information makes it impossible

to construct a timetable using the information provided from static models. On the other hand, if the frequency is used to derive timetables, the timetable is likely to be infeasible due to the conflict of trains in junctions.

The above is also supported by the International Union of Railway (UIC) which argues that "Capacity as such does not exist. Railway infrastructure capacity depends on the way it is utilised" [87]. UIC in Code 405 has proposed a methodology to assess railway capacity by evaluating line sections to identify bottlenecks [1]. The formula in Code 405 for estimating capacity is given by Equation 2.2.

$$L = \frac{T}{t_{fm} + t_r + t_{zu}} \quad (2.2)$$

In the formulation, L refers to the total capacity the line section can support measured as the total number of trains within time interval T . The term t_{fm} refers to the average time span at minimal sequence of trains (i.e. the average minimum time headway when trains are moving at average speed), t_r is the average buffer time and t_{zu} the time supplements [1]. All the values in the denominator are dimensioned as average time per train. The UIC 405 Code was succeeded by the UIC Code 406 which, rather than measuring capacity as the total number of trains, the term capacity utilisation is used instead [87]. The parameters that influence capacity utilisation are: average speed, the number of trains, stability (i.e. margins and buffers) and train heterogeneity (Figure 2.1).

Figure 2.1 illustrates the different characteristics of mixed-train timetables and

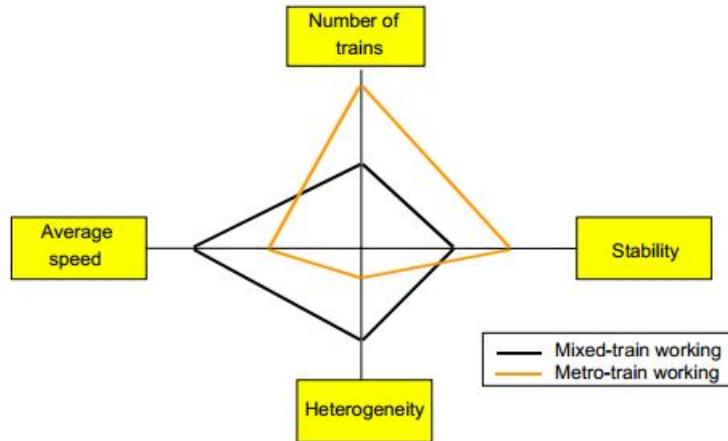


FIGURE 2.1: Railway capacity parameters [87]

metro-train timetable. For example, mixed-train timetables are comprised of highly heterogeneous services which operate at high average speeds and, in comparison to metro-services, mixed-train timetables have a smaller number of trains and are less stable. In the leaflet, UIC also presents a formula which can be used to evaluate the utilisation of network components [87]. Equation 2.3 shows the infrastructure utilisation formula proposed by UIC [87]:

$$K = \frac{A + B + C + D}{T} * 100 \quad (2.3)$$

In Equation 2.3, K is the infrastructure percentage utilisation, A the total occupation time, B total buffer time while C and D the supplements for single-track lines and supplements for maintenance respectively [87].

System capacity

System capacity has an inverse relationship with train loading and is the performance metric used by the ORR to reduce crowdedness [29]. Hence, as the system capacity of a network increases, train loading levels decrease which translates into a decrease in the levels of crowdedness. Naturally, this metric is of interest to railway passengers who are likely to feel that railways are not a satisfactory substitute to other means of transport. Subsequently, this metric may be of relevance to the government as well.

Evaluating train loading is vital for train/freight operators since it is in the operator's best interest to run the trains close to maximum carrying capacity. On the other hand, railway passengers and the government are more likely to prefer services which are less crowded. Train loading calculations require knowledge of the number of passengers in the train at any point in time and the maximum seating capacity of each train in the timetable. Albrecht [4] considers a similar cost function by estimating what is defined as operational efficiency; that is the ratio between demand and supply for passenger kilometres per unit of time. However, the value of operational efficiency calculated is an average for a given unit of time and as such information about the occupancy rate of each individual train is diluted.

Provided that an Origin-Destination matrix (O-D matrix) is made available, the

number of passengers in the train can be found through the simple recursive formula [37]:

$$P_{k(m+1)} = P_{k(m)} - \alpha_{k(m)} + \beta_{k(m)} \quad (2.4)$$

In the formula above, $P_{k(m)}$ represents the number of passengers in train k when departing from station m , $\alpha_{k(m)}$ the number of passengers alighting train k at station m and $\beta_{k(m)}$ the number of passengers boarding train k at station m .

2.2.2.2 Journey time

The importance of journey time cannot be underestimated in public transport since it is an important factor influencing the attractiveness of a means of transit [37]. In this context, journey time is defined as in-vehicle time (IVT) since other forms of journey time are either calculated in other cost functions (e.g. headway time, delay time) or are irrelevant to timetable optimisation and as such not calculated at all (e.g. walking time to the station). The importance of maintaining as low journey times as possible is well documented by numerous authors such as Mackie et al. [51, 52] and Wardman [90]. In particular, Wardman [90, 91] identifies that railway customers value their time higher than car users, highlighting the importance of achieving low journey times to maintain the competitiveness of railways. Consequently, this metric is of relevance to the regulators as well as the passengers and freight customers.

The minimisation of journey time is analysed by a number of authors [5, 11, 83]. However, these papers focus on the minimisation of travelling time without considering how many passengers are aboard. Ghoseiri et al. [37] minimise journey time while also considering the number of passengers on board but, even though they refer to the value of time concept, the actual value is not used in their formulation to express journey time as a monetary cost.

Dorfman and Medanic [32] and Li et al. [48] consider an alternative objective for the minimisation of the total travelling time whereby they try to minimise the total time J needed for all trains in the schedule to clear the line and the formulation is

$$J = t_{N_a} - t_{1_d} \quad (2.5)$$

where t_{N_a} is the arrival time of the last train in the schedule at the last node in its path and t_{1_d} the departure time of the first train in the schedule from the first node in its path. This objective is also known as the timetable's makespan (or simply span) and is also used by D'Ariano [26].

2.2.2.3 Punctuality

Punctuality modelling

Punctuality of service (defined as the extent to which trains arrive at stopping stations on time) is a performance measure highly valued by railway customers, regulators and train operators [16, 91].

Real-time railway operations are stochastic in nature meaning that operations (e.g. sectional running times and dwell times) are not constant since they are subject to disturbances which cause delays [47]. The delay of a train is taken to be the difference between its scheduled arrival time and its actual arrival time. A positive delay means that the actual arrival of a train occurs later than its scheduled arrival time, while a negative delay means that a train arrives earlier than scheduled.

Kroon et al.[74] provide three measures for assessing a timetable's robustness: primary delays that can be absorbed before they lead to knock-on delays, minimal knock-on delays from one train to the next and the ability to eliminate delays quickly. Primary delays are caused by external stochastic disturbances (i.e. any event other than the conflict with a delayed train) while knock-on delays occur when a delayed train knocks its delay on to other trains [47]. Allowance times are inserted in the timetable to absorb primary delays while buffer times are inserted to prevent the propagation of delays on to other trains.

Primary delays are modelled by fitting a probability distribution to consider the

likelihood of an external event taking place (e.g. driving behaviour differences, adverse weather conditions etc.) as well as their magnitude. Primary delays are often modelled using the exponential distribution [18, 88, 89].

Modelling delay propagation is much more complicated though than the modelling of primary delays but is still essential in evaluating the robustness of the timetable [16]. Meester and Muns [54] identify three different methods for modelling delay propagation: queuing models, analytical models and simulation models. Queueing models are generic mathematical models which are timetable independent, making them inappropriate for analysing a timetable's performance in terms of punctuality [54]. Furthermore, queueing models tend to become less accurate as the network becomes more complex, limiting their ability to provide accurate results for decision making purposes [82]. Analytical models rely on the use of conditional probability distributions (e.g. [54, 95, 96]) or heuristic approximations (e.g. [16, 18]) to incorporate knock-on delays. Such methods though require a deep understanding of statistics or are heuristics which try to approximate the effect of delays, limiting their applicability in a real world context. Robust optimisation techniques can be used to model the effects of primary and knock-on delays but the inherent conservatism of such techniques makes them inefficient for industry applications [7, 8, 36, 45]. Lastly, simulation models can take too long to run, limiting their applicability [16, 54]. Consequently, since all methods have their limitations, the purpose of modelling delays should be carefully taken into account in order to determine which one of the above methods is the most appropriate.

Measuring punctuality

Over the years, a number of different methodologies have been developed aimed at evaluating the punctuality of a railway timetable. Meester and Muns [54] propose three methods for calculating penalties associated with delays:

- Expected fraction of arrivals at most n minutes late
- Average expected delays
- Average expected penalty on delays above n minutes

In their optimisation model though, Dorfman and Medanic [32] and Li et al. [48] use three different performance measures to assess a timetable's reliability:

- Total delay experienced by all trains
- Maximum delay experienced by a train
- Timetable time-efficiency

The formulation for the timetable time-efficiency objective is:

$$\eta = \frac{t_{N_a}^f - t_{1_d}^f}{t_{N_a}^{ob} - t_{1_d}^f} \quad (2.6)$$

where $t_{N_a}^f$ is the scheduled time of arrival of the last train in the schedule, $t_{1_d}^f$ the scheduled departure of the first train in the schedule and $t_{N_a}^{ob}$ the actual arrival

of the last train in the schedule. As the timetable experience delays from which it cannot recover, $t_{N_a}^{ob}$ will increase, making the denominator larger which in turn provides lower values for the timetable's time-efficiency with respect to reliability of service. Goverde et al. [40] and Sama et al. [80] use a wider range of objectives to calculate delays which have all been listed in Section 2.2.1.

However, the above two lists are not conclusive and, in literature, a number of different methods are being utilised to analyse the reliability of a given timetable. For example, Vansteenwegen and Van Oudheusden [88, 89] minimise a function which penalises weighted waiting times which result from primary delays. Peterson [74] minimises primary delays of two services in the timetable by redistributing the allowance times in the existing timetable. He also uses different weights to reflect the fact that passengers weigh delay time higher than travel time [74]. Carey and Kwiecinsky [18] minimise total primary and secondary delays by inserting buffer times but their problem is very small in size. Liebchen et al. [49] focus on delay resistant timetables but only transfer passengers are considered when evaluating a timetable's performance. Kraay and Harker [46] present a scheduling formulation which aims to reduce delays but their focus is only on freight trains and their objective function is divided into two parts which are being minimised simultaneously. The first term penalises actual arrival and departure time deviations from the scheduled time, while the second part of the objective function penalises missing scheduled connections and violations of the 12 hour rule (the maximum number of driving hours).

Finally, it should be noted that, in the United Kingdom, punctuality is measured as the percentage of services which arrive on time at their terminal stations. This punctuality metric is widely known as Public Performance Measure (PPM) [60]. Commuter services which arrive within 5 minutes from their scheduling time are assumed to arrive on time while that number rises to 10 minutes for long distance services [60]. This measure of performance though has some serious disadvantages as it does not consider other stations in the train's path while it also does not provide information on how late a train is. For example, a commuter service which is 6 minutes late is treated in the same way as a train which is 30 minutes late, undermining the usefulness of this performance measure. The rail industry also measures delays using what is known as 'delay minutes' which are defined as '...a loss of time against a schedule between two consecutive locations on the train's journeys' [59]. This metric is currently being used to determine the penalty that Network Rail or the Train Operator have to pay (depending on who will be allocated responsibility for the delay) as a result of the delay [66]. Consequently, a train which is described as 'on-time' using the PPM, may have accumulated 'delay minutes' on its way until the terminal station.

2.2.2.4 Waiting time

Waiting time is found by estimating the time customers have to spend on the platform waiting for their service to arrive. Albrecht [4] minimises the mean waiting time of passengers using the average wait formulation by Osuna and Newell

[70]. The formulation considers the headway between services and, by assuming that customer arrivals are uniformly distributed, the expected average wait is estimated. Calculating the average though has the downside that if a group of customers wait for too long, the impact of their waiting time can be mitigated if the rest of the waiting times are short enough.

Albrecht notes that suburban trains have easily recallable departure times and this, in conjunction with the availability of pre-trip and on-trip information, allows for passengers to arrive at the platform just in time to catch the train [4]. However, the fact that passengers are aware of the scheduled departure time of their service does not imply that demand for a service does not exist; it may as well exist but not being served frequently due to the sporadic arrivals of the service. The importance of this performance metric is identified by Wardman [91] who states that: 'Public transport users can either plan their activities around scheduled departure times, which involves inconvenience and transaction costs along with some amount of wait time, or else turn up at the departure point at random, which avoids the scheduling costs but incurs additional waiting...'. This is an idea also shared by the Department for Transport [78] which claims: '...the time people actually spend waiting at a station or stop might not fully reflect the inconvenience of the service frequency, which might also affect when people have to (rather than when they would prefer to) leave or arrive'. Therefore, it is preferable to have regular headways between services while also communicating that information to the public to prevent them from experiencing any inconveniences related to excessive waiting.

2.3 Optimisation algorithms

The train timetabling problem is an NP-Hard problem³ and, as a consequence, good heuristics and meta-heuristics are necessary to obtain solutions which are close to optimality [13, 17, 32, 42, 68]. At the moment, a number of different optimisation algorithms are being used which can output a feasible timetable. To eradicate the problem faced by exact algorithms, multiple heuristics have been applied over the years to tackle the train scheduling problem. Unlike exact algorithms, heuristics attempt to find approximate solutions to the problem within a reasonable period of time.

Several papers propose exact algorithms to solve the problem by implementing variations of the Branch and Bound algorithm [39, 42, 57]. However, due to the computational complexity of the problem, the efficiency of such algorithms suffers severely when the problem grows in size. Branch and Bound algorithms can be used as a heuristic by terminating them before they converge to the global optimum [26]. Branch and bound algorithms can be terminated before reaching the optimal so they rely on their ability to converge to a good solution within an acceptable time interval. One of the most recent and used algorithms for tackling timetabling problem is the Branch and Bound algorithm developed by D'Ariano which formulates the problem as an alternative graph [23, 26]. The problem then becomes a job-shop problem with no store⁴ as well as the constraints relevant to

³Optimisation problems classified as NP-Hard are those problems for which no polynomial algorithms exist that can solve the problem to optimality [71]

⁴The job-shop problem is a class of problems where a number of jobs need to be processed by one or more resources (also known as machines) with each resource needing a given amount

timetable optimisation. A branch and bound procedure takes advantage of specific problem characteristics so that the algorithm can be truncated relatively quickly as it converges to good solutions in very little time [19, 26, 27].

A constraint generation algorithm is proposed by Odijk [68] specifically designed to solve the periodic timetabling problem. The algorithm formulates constraints which capture the periodic time window of each timetable (called timetable structure) and uses a branch and bound in conjunction with a feasible differential algorithm to determine whether a feasible solution to the periodic timetable problem exists.

Genetic algorithms are a well known class of heuristics which numerous authors have relied on in the past to approach large scale timetabling problems [5, 58, 83]. Each implementation differs in terms of the problem encoding and the way the timetable is determined. For example, Suttewong [83] makes the use of two types of binary variables to encode the problem, the first variable is encoded as a three dimensional array and the entry $x_{i,j,s}$ takes a value of one if train i visits node s before train j . The second variable is encoded as a two dimensional array and the entry $Y_{i,s}$ takes the value of one if a train i utilises node s and is zero otherwise.

On the other hand, Barber et al. [5] encode the problem using a single binary variable which contains information about the sequence with which all trains will

of time to process each job. The goal of the problem is to find a way to schedule each job to each machine such that the timespan of all the jobs is minimised [19]. The 'no store' variation prevents each resource from storing a job and accepting another one before passing on the job it processed. This means that once a resource starts processing a job, the job needs to move to the next resource before the current resource accepts any new jobs [19]

visit the stations in their path. The departure time from a given node is calculated by finding the closest feasible node.

Nonetheless, several other optimisation techniques are being used such as sub-gradient optimisation algorithms, greedy heuristics, simulated annealing and Lagrangian relaxation heuristics [14, 37]. Finally, certain researchers rely on the use of simulation to find approximate solutions to the problem [32, 48].

2.4 Summary

At the moment, even though literature exists on the optimisation of railway timetables using a wide range of objectives, when it comes to the simultaneous optimisation of numerous objectives, literature is quite limited. This is because the vast majority of the authors only consider two objectives simultaneously and these objectives are usually shaped according to the needs of a single railway stakeholder. This might also explain the lack of research on how to find a common dimension to measure numerous timetabling objectives.

The reason for only choosing two objectives lies in the fact that researchers, quite often, want to analyse the trade-off between cost functions so they only pick two objectives to prevent the impact of a third objective interfering with the results. Authors defend their decision to use only two objectives by arguing that in the timetabling process, capacity and punctuality are the main metrics of interest

while for real-time rescheduling, punctuality and energy consumption are the main metrics of interest.

With regards to network capacity, static capacity estimation approaches estimate the maximum number of trains which can traverse the network in a given amount of time but the static nature of such approaches makes them inappropriate for timetabling. UIC Codes 405 and 406 offers different methods of estimating capacity which both rely on a given timetable. Code 405 counts the total number of trains a network can support while Code 406 measures infrastructure utilisation instead. System capacity (or train loading) is an objective which is not commonly used in optimisation and the only paper found to consider it, measures the average utilisation of all trains without taking into account the carrying capacity of each train as well as the number of passengers in it. Journey time is commonly used for optimisation due to its importance for numerous stakeholders but few authors compute journey time by considering the number of passengers on board. Punctuality is an objective used very often in timetabling optimisation and there is extensive literature on how to model primary and knock-on delays as well as how to penalise delays. We will be measuring waiting time by estimating the time customers spend on the platform waiting for the train to arrive. This objective has not been extensively studied in literature and the only paper found to use waiting time at the platform for multi-objective optimisation, uses it in conjunction with average train loading. The above is not a conclusive list of performance metrics used in timetabling and a more detailed discussion of other performance metrics can be found in [20] and [77].

Finally, numerous optimisation algorithms are in place for solving the train timetabling problem but, due to the computational complexity of the problem, exact algorithms become inefficient as the problem grows in size. Therefore, multiple heuristics have been developed to find approximate solutions to the problem while some authors resort to the use of simulations.

Chapter 3

Cost Functions

3.1 Introduction

This chapter explains the methodology used to formulate the cost functions and the method used to make sure that all cost functions have the same dimension. The cost functions measure a wide range of performance metrics which may concern multiple stakeholders in the railway industry. The cost functions are subsequently re-dimensioned such that they measure a timetable's monetary cost; a metric easily understood by both academics and industry professionals.

Section 3.2 introduces the different variables used in formulating the cost functions and explains how they relate to the train timetabling problem. Section 3.3 provides the formulation of the cost functions and finally Section 3.4 explains how the

concept of travel time savings (also known as values of time) is employed so as to calculate the monetary cost of each individual cost function.

3.2 Specification of timetable and associated constraints

A timetable is typically constructed by specifying the arrival $\tau_{n,i}$ and departure times $\sigma_{n,i \rightarrow j}$ of each train n over a set of control points i, j ($\forall i \neq j$) (which can be a station, junction, etc.) along its service route. An example is shown in Figure 3.1 in which the horizontal and vertical axes represent the time and position along the train route respectively. Each line on the diagram represents a train run which is specified by a series of departure $\sigma_{n,i \rightarrow j}$ and arrival times $\tau_{n,i}$ at station i for each train n as specified by the timetable. Given a set of $\sigma_{n,i \rightarrow j}$ and $\tau_{n,i}$, we can derive the running time $T_{n,i \rightarrow j}$ of each train n between station i and j as

$$T_{n,i \rightarrow j} = \tau_{n,j} - \sigma_{n,i \rightarrow j}, \quad (3.1)$$

and also the dwell time $D_{n,i}$ of train n at station i

$$D_{n,i} = \sigma_{n,i \rightarrow j} - \tau_{n,i}, \quad (3.2)$$

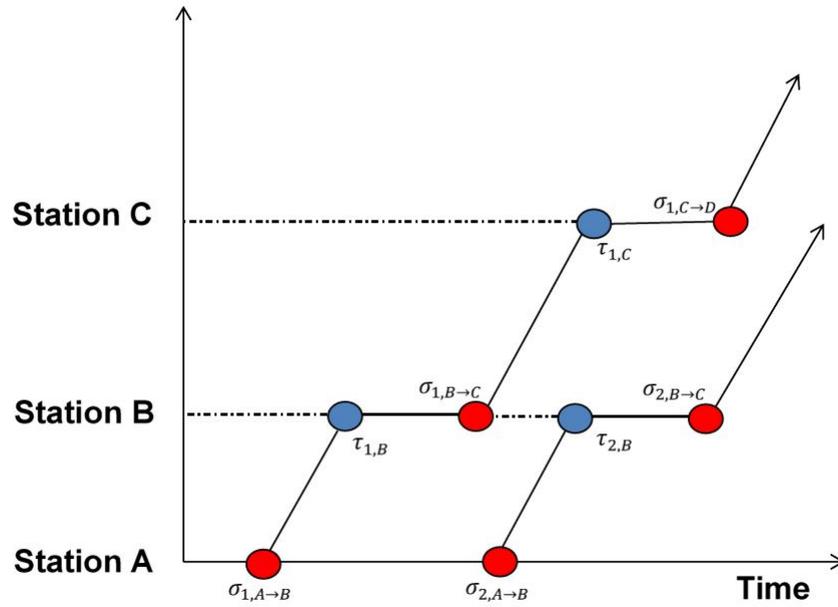


FIGURE 3.1: Timetabling variables illustration

The setting of the variables $\sigma_{n,i \rightarrow j}$ and $\tau_{n,i}$ will be subject to a set of operational constraints in practice. We first have the minimum sectional running time constraints to reflect the speed limit imposed on each track section:

$$\tau_{n,j} \geq \sigma_{n,i \rightarrow j} + \frac{\Delta_{i,j}}{v_n^*}, \quad (3.3)$$

where $\Delta_{i,j}$ is the distance between stations i and j , v_n^* is a constant representing the maximum speed a train can achieve given the train's maximum speed and the speed limit on the current track section for train n travelling from station i to j . This means that, for the purpose of this project, a train's motion in any given section will not be modelled using any dynamic information such as acceleration and deceleration. The formulation only provides a lower bound for the time needed for a train to travel any given distance. The exact method for

calculating the running times of trains is given in Section 4.5.4. Moreover, we also have the minimum dwell time constraints which define the minimum time each train n has to spend at station i :

$$\sigma_{n,i \rightarrow j} - \tau_{n,i} \geq D_{n,i}^*, \quad (3.4)$$

The minimum dwell time $D_{n,i}^*$ imposed here will typically be determined by a number of factors on the demand side such as demand level of passengers or freight for that specific train at that specific station, and/or the consideration of connectivity where it is necessary to ensure a long enough dwell time for passengers or goods to transfer from one train to another at the station or interchange [73].

To implement the signalling system, each track section is further disaggregated into a series of blocks. Under the current fixed block signalling systems in practice, each block can only accommodate up to one train at a time to ensure safe operations (see Figure 3.2). Referring to Figure 3.2, denote the arrival and departure times of train n at block b between station pair (A, B) as $\tau_{n,b}$ and $\sigma_{n,b \rightarrow B}$ respectively. The shaded region in the figure represents the location and time period (during times $t_{n,b}^{in}$ and $t_{n,b}^{out}$) that is occupied by the train of interest during which other trains are prohibited from entering. Following the specification in the current UIC (International Union of Railways) operational code [87], we have:

$$t_{n,b}^{in} = \tau_{n,b} + \frac{\delta_{n,b}}{v_{n,b}}, \quad (3.5)$$

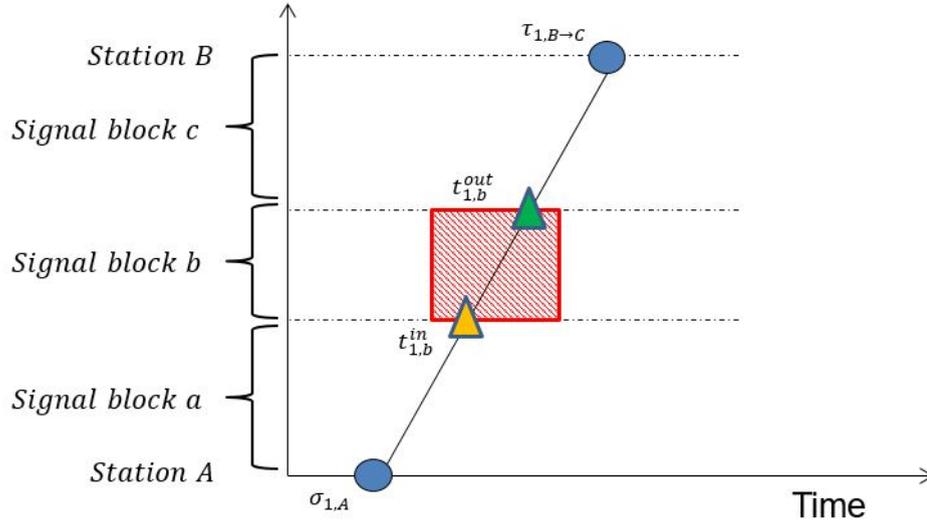


FIGURE 3.2: Signalling block example occurrence

where $\delta_{n,b}$ is the visual distance of train n to the entrance of block b ; $v_{n,b}$ is the nominal speed of train n travelling through block b . The time $t_{n,b}^{in}$ represents the time when the driver of train n observes the signal aspect at block b and starts to take according action(s). Moreover, the time at which block b is released from train n is defined as:

$$t_{n,b}^{out} = \sigma_{n,b \rightarrow c} + \frac{L_n}{v_{n,b}}, \quad (3.6)$$

where L_n is the length of train n . The time $t_{n,b}^{out}$ represents the time when the tail of the train n clears from the block section b and enters block section c . Because of the signalling system, congestion is expected to occur when the train volume on a track section is high [31, 38]. Following the definitions of $t_{n,b}^{in}$ and $t_{n,b}^{out}$ set in Equations 3.5 and 3.6 respectively, the signal blocking constraint can then be

written mathematically for all station pairs (i, j) and signal blocks b as

$$\tau_{n+1,b} \geq \sigma_{n,b \rightarrow c} + \frac{L_n}{v_{n,b}}, \quad (3.7)$$

in which train $n + 1$ is the train following immediately after train n in block section b . This constraint prevents trains from simultaneously occupying a signalling block. Finally, a headway constraint is imposed which maintain safety time margins between trains. This constraint is formulated as

$$\tau_{n+1,s,j} \geq \tau_{n,i} + h_b^* \quad (3.8)$$

where h_b^* denotes the minimum time headway which must be kept between the arrival time of two trains at any time in signalling block b .

A detailed formulation of the train scheduling problem is given by multiple authors such as Ghoseiri et al. [37], Higgins et al. [42] and Barber et al. [5]. However, the constraints identified above, in conjunction with the optimisation procedure outlined in Chapter 5, ensure that feasible timetables are generated which can be evaluated using the cost functions formulated in Section 3.3.

3.3 Performance metrics and cost functions

With the timetable and the associated constraints specified, we can then formulate the cost functions to be used in the optimisation framework. Following the

comprehensive review in [20] and [77], we have selected four representative performance metrics in the railway industry: train running times, customer waiting times, service punctuality and crowdedness. We expect that all cost functions are of interest for train operators and passengers while the Government is likely to be interested in monitoring the performance of the last three metrics although the Government might want to have in mind the journey times to make sure that railways remain competitive. Finally, the Infrastructure Manager will be more interested in punctuality of services.

3.3.1 Running times of trains

The running times ($T_{n,i \rightarrow j}$) of trains n over all sections (i, j) can be obtained from Equation 3.1 in the previous section following the specification of timetable variables $\sigma_{n,i \rightarrow j}$ and $\tau_{n,i}$. Given all running times $T_{n,i \rightarrow j}$, we define the cost associated with the running time components as:

$$C_T = c_T \sum_{n=1}^N \sum_{\{(i,j) \in S_n\}} T_{n,i \rightarrow j} p_{n,i}, \quad (3.9)$$

where N represent the total number of trains and S_n the stations in the path of train n . The variable $p_{n,i}$ is a quantity associated with the passenger demand for train service n running between stations i and j . With this $p_{n,i}$, the corresponding timetable will then give higher priority to trains carrying more passengers. Finally, the notation c_T represents a monetary cost associated with the running times. We

will have further discussion on the choice of this c_T and other monetary cost coefficients in Section 3.4.

3.3.2 Waiting times of passengers

The waiting time cost function will penalise the time spent by passengers waiting for their service to arrive. Estimating the cost associated with waiting times first requires knowledge of $\lambda_{i \rightarrow j}(t)$ which denotes the demand profile for passengers requesting a service from station i to station j over time t . Fundamental queueing analysis (e.g. [25]) gives the total waiting time W (in the unit of [persons-time]) as

$$W = \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \iint_{\tau_{n,i}}^{\tau_{n+1,i}} \lambda_{i \rightarrow j}(t) dt^2, \quad (3.10)$$

where N_s is the total number of trains serving station s over the study time period. The nested summation over the elements $\{n \in (N_i \cap N_j)\}$ serves the purpose of prohibiting passengers from boarding trains which do not stop in the stations the passengers demand. Consequently, the set intersection makes sure that train n stops both at station i and station j . The time interval between $\tau_{n,i}$ and $\tau_{n+1,i}$ specifies the headway of train service at station i which will also serve station j . Equation 3.10 can be simplified by assuming a uniform demand $\bar{\lambda}_{i \rightarrow j} = \lambda_{i \rightarrow j}(t)$

for all times t during the study period as:

$$\begin{aligned}
W &= \sum_{\{\forall(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \iint_{\tau_{n,i}}^{\tau_{n+1,i}} \bar{\lambda}_{i \rightarrow j} dt^2, \\
&= \sum_{\{\forall(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \int_{\tau_{n,i}}^{\tau_{n+1,i}} \bar{\lambda}_{i \rightarrow j} (\tau_{n+1,i} - \tau_{n,i}) dt, \\
&= \sum_{\{\forall(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \bar{\lambda}_{i \rightarrow j} [(\tau_{n+1,i} \tau_{n+1,i} - \tau_{n,i} \tau_{n+1,i}) - (\tau_{n,i} \tau_{n+1,i} + \tau_{n,i} \tau_{n,i})],
\end{aligned}$$

$$\therefore W = \sum_{\{\forall(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \bar{\lambda}_{i \rightarrow j} [\tau_{n+1,i} - \tau_{n,i}]^2. \quad (3.11)$$

As reflected from Equation 3.11, the total waiting time grows linearly with the average demand rate $\bar{\lambda}_{i \rightarrow j}$ but quadratically as the service headway increases (i.e. frequency of service decreases). However, the uniform demand assumption made in deriving Equation 3.11 may be valid for high frequency service (e.g. metro) while it may not be appropriate for low frequency mainline services as it is known that the arrival of passengers will cluster around the publicised scheduled service times in the timetable. Hence some detailed survey will be needed for obtaining the demand pattern if one wants to have a reasonable estimate of waiting times when deriving mainline timetable.

Finally, following the calculation of W , the eventual cost associated with waiting times is determined as:

$$C_W = c_W \sum_{\{\forall(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \iint_{\tau_{n,i}}^{\tau_{n+1,i}} \lambda_{i \rightarrow j}(t) dt^2, \quad (3.12)$$

where \hat{c}_W is the monetary cost associated with waiting times.

The purpose of incorporating the waiting time into the optimisation framework is to ensure that there are enough services for number of passengers or goods at the station without creating excessive waiting times. Empirical studies conducted by the UK Department for Transport (e.g. [85, 90, 91]) suggest that this c_W will be around two or three times larger than c_T as the waiting time is generally regarded as a dead loss. More information can be found in Section 3.4.

3.3.3 Punctuality of service

Punctuality is measured herein as the time discrepancy between the scheduled and the actual arrival times of the train services. To quantify the punctuality in monetary units (see [15, 66]), we adopt a punctuality cost function as shown in Figure 3.3.

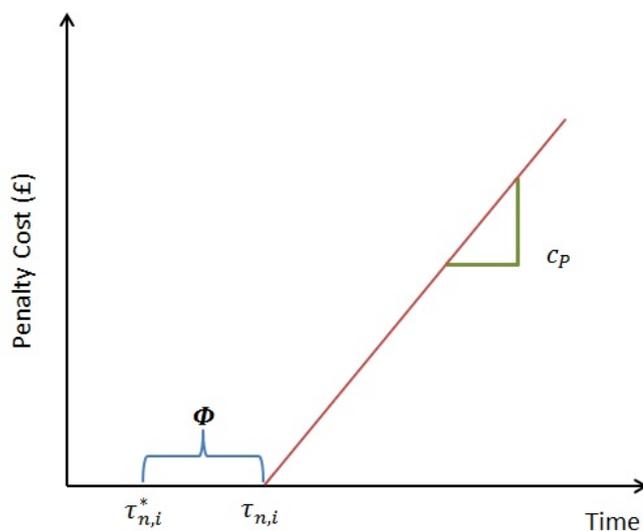


FIGURE 3.3: Punctuality cost function

In the figure, $\tau_{n,i}^*$ denotes the scheduled arrival time of the train service while Φ is a time allowance for lateness meaning that no time penalty is charged if the train arrives within its allowance time (e.g. Φ is considered to be three minutes under the UK railway operational regulations [66]). If the corresponding train is delayed by more than Φ from the scheduled arrival time $\tau_{n,i}^*$, a schedule delay cost will be imposed on the Train Operator by the Infrastructure Manager for lateness. It is considered here that this schedule delay cost increases linearly with a slope of c_P over arrival time $\tau_{n,i}$, where $\tau_{n,i} \geq \tau_{n,i}^* + \Phi$. This penalty rate c_P represents the loss in value of time of customers (passengers or freight companies) per unit lateness in time [66, 69]. Following this linear specification, the total schedule delay cost associated with punctuality can be determined, taking the arrival of passengers and/or goods into account, as

$$C_P = c_P \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \int_{\tau_{n,j}}^{\tau_{n+1,j}} \lambda_{i \rightarrow j}(t) (\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+ dt, \quad (3.13)$$

where $\tau_{n+1,i}^*$ is the arrival time for train $n+1$ at station s as given in the timetable, $(\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+ = \max[(\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi), 0]$. Similar to Equation 3.10, Equation 3.13 can be simplified by assuming uniform arrival $\bar{\lambda}_i = \lambda_i(t)$ for all times t as

$$\begin{aligned} P &= \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \int_{\tau_{n,j}}^{\tau_{n+1,j}} \bar{\lambda}_{i \rightarrow j} (\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+ dt, \\ &= \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \bar{\lambda}_{i \rightarrow j} (\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+ \left[t \right]_{\tau_{n,j}}^{\tau_{n+1,j}}, \end{aligned}$$

$$\therefore P = \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \bar{\lambda}_{i \rightarrow j} (\tau_{n+1,j} - \tau_{n,j}) (\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+. \quad (3.14)$$

Finally, it is noted that this punctuality cost analysis is generally applicable to other schedule cost functions, apart from the linear assumption in Figure 3.13, by revising the cost function term

$$(\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+ \quad (3.15)$$

in Equation 3.13.

Certain authors (e.g. [88, 89]) also penalise trains when arriving at a station ahead of schedule. The rationale for penalising early arrivals is that passengers who will not exit at the current station, will incur a penalty for waiting rather than travelling. The penalty for waiting is higher than the penalty for travelling so early arrivals are also penalised. The British Department for Transport defines waiting time as the time that passengers spend on the platform waiting for their service [52, 85]. Since the definition of waiting time provided by the Department of Transport does not incorporate the in-vehicle waiting time penalised by Vansteenwegen and Van Oudheusden [88, 89], no penalty will be applied in the case of early arrivals. Another justification for not penalising early arrivals is the fact that if a train arrives early it may impact other trains leading to their delay and this delay will be captured by the cost function provided in Equation 3.13. If on the other hand a train's early arrival does not impact on other trains, then there is no harm in arriving at a station early.

3.3.4 Crowdedness

The crowdedness cost function measures the difference between travelling in crowded versus uncrowded trains and is an important aspect for the passengers when assessing the attractiveness of public transport [92]. The difference is derived from the fact that passengers value their time higher when they travel in crowded trains.

$$C_D = c_D \sum_{n=1}^N \sum_{\{(i,j) \in S_n\}} R_{n,i \rightarrow j}(p) p_{n,i} T_{n,i \rightarrow j} \quad (3.16)$$

In the above formulation, R denotes the time multiplier, given the number of passengers on board, relative to the train's seating capacity. At low crowdedness values, the time multiplier is equal to zero meaning that no penalty is charged for overcrowding. After the train's loading levels exceed a given threshold, the penalty increases linearly with the train loading [86, 92]. The time multiplier for standing passengers is much higher than that for seating passengers to reflect the increased dissatisfaction of passengers when they are unable to find a seat [86, 92].

3.4 The cost of travel time savings

The history of travel time savings (also known as value of time) in the UK starts in the 1960s with the need to evaluate a journey's non-monetary costs to carry out cost benefit analysis for the construction of the M1 motorway and the Victoria

Line in London [52]. The Department for Transport classifies travel-related costs into two broad categories [52, 78].

- Monetary costs cover the travel-related costs that a person must pay using real life currency (e.g. the cost of purchasing a train ticket, the cost of refuelling the vehicle).
- Non-monetary costs are being used to penalise a wide variety travelling behaviours such as in-vehicle time, waiting time and walking time. These costs do not involve the exchange of real-life currency so their monetary value is estimated by monetising the passengers' time.

Non-monetary costs along with any monetary costs comprise what is known as the generalised cost of a journey and represent the opportunity cost (in financial terms) of travelling [52].

Over the following years, it became evident that different time valuations should be calculated depending on whether the passenger is travelling during working hours or not. Consequently, time valuations were estimated for travelling during working and non-working time [52]. Subsequent research [51, 52, 85, 90, 91] identified three different passenger types

- Business passengers are the passengers who travel during working hours
- Commute passengers travel to and from work

- Leisure passengers travel for any other purpose except the two mentioned above

Recent research findings though show that the time valuations for commuting and business passengers travelling a short distance are close to each other, blurring the lines between the time valuations for different journey purposes [78].

The travel time for business passengers is valued differently depending on the mode they are travelling while for commute and leisure passengers their time valuation is mode-independent [85]. The reason for the business values being mode-dependant is because the values are based on the average income of business passengers using each specific mode [52, 78, 90, 91]. Business passengers travelling via rail were found to have the highest VoTs followed by car and bus passengers [78, 85].

Recently, VoTs has been used by the DfT on a strategic level to carry out a cost-benefit analysis to evaluate the impact of transport investments such as Crossrail and the High Speed Two (HS2) lines [9, 43, 78]. The fact that VoTs are used to evaluate such important and expensive projects shows the importance the British government places on evaluating a project's non-monetary costs using travel time savings.

The fact that all cost functions measure passenger hours enables us to apply the 'value of time' concept in order to transform all cost functions such that they are expressed in monetary terms. The monetary coefficients in the cost functions are set from official documents published by the British Department for Transport and

Network Rail. The monetary coefficients are set according the 'webTAG Unit 3.5.6' guidance [85] published by UK Department for Transport which specifies the values of time of travellers based on an empirical study conducted by University of Leeds [52]. Due to confidentiality issues outlined in Section 3.5, calculating a timetable's monetary cost is not possible. Therefore, by applying time valuations to our cost function formulations, it is possible to assess a timetable's non-monetary cost.

Valuations of journey time differ according to the passenger's travel purpose so different time valuations are given depending on each purpose. Furthermore, the time for waiting, arriving late and travelling in crowded trains is given by multiplying the passengers' travelling time by a time multiplier which represents the opportunity cost of the passenger for waiting, arriving late and travelling in crowded trains. Sections 3.4.1 to 3.4.3 give the monetary coefficients to be applied to each performance metric while Section 3.4.4 updates formulation of the cost functions.

3.4.1 Values of travelling and waiting

The time valuations for railway passengers are provided in the 'Passenger Demand Forecasting Handbook' and the validity of the VoTs for railway passengers has been further enhanced by research carried out by the DfT in 2015 [78, 85]. An analysis carried out in recent years by the DfT [78] was designed to understand whether the VoTs for business passengers using railways should be adjusted to reflect the fact that passengers are now able to use mobile devices and access the internet

while on board. However, the research failed to report any statistically significant changes in the VoTs due to technological developments [78].

Following the guidelines set in WebTag 3.6.5 and the Passenger 'Demand Forecasting Handbook', the cost of waiting is set to 2.5 times the cost of travelling [78, 85]. Research carried out by the DfT in 2015 has shown that the waiting time multiplier should be reduced to 2.0 but, as of the time of writing, this revision has not been made official [78].

The final costs used in this project to penalise travelling time and waiting time are summarised in Table 3.1. These values are officially used by the DfT in WebTag 3.6.5 and the 'Passenger Demand Forecasting Handbook' [78, 85]. The meaning of the costs can be interpreted to be the opportunity cost of travelling in financial terms.

TABLE 3.1: Monetary coefficients [85]

Cost Function	Coefficient	Value of time of each passenger type (£/hour)		
		Business	Commute	Leisure
Journey Time	C_T	£31.96	£6.81	£6.04
Waiting Time	C_W	£79.90	£17.03	£15.10

3.4.2 Punctuality multipliers

Train delays are being penalised through the use of punctuality multipliers which are applied to the value of time [78]. Punctuality multipliers differ for each mode

so the multipliers for lateness are taken from the recommended values included in the 'Passenger Demand Forecasting Handbook' [78]. Research carried out in 2015 shows that the value of punctuality multipliers has not changed significantly since the publication of the 'Passenger Demand Forecasting Handbook' [78]. Table 3.2 illustrates the values of the punctuality multipliers used in the project. Unlike the

TABLE 3.2: Lateness multipliers [78]

Flow type	Less than 20 miles		More than 20 miles	
	Commuting	Other	Commuting	Other
London TCA ^a	2.5	2.3	2.5	2.3
SE ^b - London	2.5	2.3	2.5	2.3
SE ^b - SE	3.0	2.3	3.9	3.4
London - Outside LSE ^c	2.5	3.0	2.5	3.0
Non LSE ^c	3.0	2.3	3.9	3.4
Airports	6.0	6.0	6.0	6.0

^a London Travel Card Area

^b South East

^c London and South East

multipliers presented in Section 3.4.1, punctuality multipliers depend on the flow type (e.g. if the train travels from London to south east) and also on distance.

3.4.3 Crowdedness multipliers

Crowdedness in trains is penalised by the DfT since it is assumed to lead to lower comfort levels for the passengers and decreased productivity while in the train [78].

The monetary coefficients for the Crowdedness cost function are the same as in the Journey Time cost function since they both penalise travelling time. The time multipliers for travelling in crowded trains have been published by Network Rail

[86] and are supported by the findings reported by a number of authors [84, 92].

In general, both the industry and academia seem to agree that the multipliers increase linearly with crowdedness levels [78, 86, 92]. Different time multipliers are applied depending on the geographical area, crowdedness levels, passenger type and whether the passengers are sitting or standing [86]. These time multipliers can be seen in Table 3.3 and Table 3.4.

TABLE 3.3: Sitting penalties for crowdedness [86]

Load Factor ^a	London-based Services					Non-London-based Services		
	Leisure	Business		Commute		Leisure	Business	Commute
		Standard	First Class	Outer	Inner			
60%	-	-	-	-	-	-	-	-
70%	0.04	0.04	-	-	-	0.02	0.04	-
80%	0.07	0.08	-	-	-	0.04	0.08	-
90%	0.15	0.16	0.23	0.04	-	0.07	0.11	0.11
100%	0.20	0.23	0.47	0.07	0.08	0.09	0.15	0.22
110%	0.27	0.31	-	0.11	0.16	0.15	0.21	0.32
120%	0.33	0.39	-	0.14	0.23	0.22	0.27	0.43
130%	0.40	0.46	-	0.18	0.31	0.28	0.34	0.54
140%	0.45	0.54	-	0.22	0.39	0.35	0.40	0.65
150%	-	-	-	-	0.47	-	-	-
160%	-	-	-	-	0.55	-	-	-

^a 'Load factor' is the percentage of passengers on board relative to a train's seating capacity.

TABLE 3.4: Standing penalties for crowdedness [86]

Load Factor ^a	London-based Services					Non-London-based Services		
	Leisure	Business		Commute		Leisure	Business	Commute
		Standard	First Class	Outer	Inner			
100%	2.12	1.70	-	1.28	1.28	2.12	2.86	1.76
110%	2.33	1.87	-	1.33	1.33	2.33	2.93	1.89
120%	2.54	2.04	-	1.38	1.38	2.54	3.01	2.03
130%	2.75	2.21	-	1.44	1.44	2.75	3.08	2.16
140%	2.96	2.38	-	1.49	1.49	2.96	3.15	2.30
150%	-	-	-	1.54	1.54	-	-	2.43
160%	-	-	-	1.60	1.60	-	-	2.57

^a 'Load factor' is the percentage of passengers on board relative to a train's seating capacity.

To calculate multiplier values up to 300%, a linear extrapolation must be carried out from the multiplier values at 120% and 140% loading factors [86].

As obvious from Tables 3.3 and 3.4, seated commuters have higher valuations of their time compared to business passengers in non-London based services. At a first glance, this seems as a counter-intuitive result since the time valuations for travelling, waiting and arriving late shows that business passengers have by far a higher valuation of their time. One potential reason which may explain this paradox is given by [55] in which it is claimed that business passengers' ability to work is not affected significantly by the levels of crowdedness. This may be due to the fact that business passengers are more likely to plan in advance and as such secure seats which favour working (e.g. table seats) [55]. It is recognised though that further analysis must be carried out so as to draw definitive inferences. The time multipliers for standing passengers are more intuitive since business passengers have the highest time valuations.

3.4.4 Cost function formulations

Following the specification of the monetary coefficients mentioned in Sections 3.4.1 to 3.4.3, the cost functions can be formulated as:

$$C_T = \sum_{y=1}^Y c_T^y \sum_{n=1}^N \sum_{\{(i,j) \in S_n\}} T_{n,i \rightarrow j} p_{n,i}^y, \quad (3.17)$$

$$C_W = \sum_{y=1}^Y c_W^y \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \iint_{\tau_{n,i}}^{\tau_{n+1,i}} \lambda_{i \rightarrow j}^y(t) dt^2, \quad (3.18)$$

$$C_P = \sum_{y=1}^Y c_P^y \sum_{\{(i,j) \in S\}} \sum_{\{\forall n \in (N_i \cap N_j)\}} \int_{\tau_{n,j}}^{\tau_{n+1,j}} \lambda_{i \rightarrow j}^y(t) (\tau_{n+1,i} - \tau_{n+1,i}^* - \Phi)^+ dt, \quad (3.19)$$

$$C_D = \sum_{y=1}^Y c_D^y \sum_{n=1}^N \sum_{\{(i,j) \in S_n\}} R_{n,i \rightarrow j}^y(p, a_n) p_{n,i}^y T_{n,i \rightarrow j} \quad (3.20)$$

where c^y specifies the monetary coefficient for each passenger type y ($\forall y \in Y$).

3.5 Summary

Section 3.2 of this chapter provides the definition of variables used for the purpose of this study and it also formulates the constraints used to construct a feasible timetable. Section 3.3 presents four performance metrics widely used in the railway industry and the formulations of the corresponding cost functions while Section 3.4 illustrates the monetary coefficients associated with each of the four cost functions.

In terms of the contribution of this project in the area of formulations of cost

functions for railway timetables, the inclusion of the crowdedness metric is a novel idea which enables the evaluation of a network's capacity. This is a significant contribution since, until now, authors evaluate the trade-off between network capacity and punctuality but their analysis does not help decision makers in deciding the number of trains to schedule. For example, Equation 2.1 suggests that at very low utilisation levels, scheduling one additional train may be a wise decision since the deterioration in timetable punctuality is minimal. However, if passenger demand is low, scheduling one additional train may not be the best option since the decrease in the cost of crowdedness is likely to be surpassed by the increase in the cost of punctuality. This is an obstacle that the formulation for the cost of crowdedness manages to overcome. With regard to the rest of the cost functions, a number of authors investigate formulations which are similar to the punctuality and journey time cost functions provided in this work while no author has been found to provide a similar formulation for calculating the waiting times. Furthermore, the fact that each cost function is evaluated in terms of its monetary cost, is not something that has been used in academia for any metrics, apart from journey time and punctuality, while in the British railway industry such an analysis is only being carried out at a strategic level. This means that no formulations are being used in the operational level to evaluate timetable related performance metrics.

It should be emphasised that the four cost functions considered herein do not present all the possible performance metrics of a timetable's performance. Other performance metrics such as train loading, track utilisation and energy consumption can be included to provide a more all-round assessment of assessing a timetable's

performance. However, no reliable monetary costs could be attributed to train loading and track utilisation due to the strict confidentiality which surrounds the cost of scheduling train services. Efforts have been made to obtain information about operating costs and the costs of the franchises but they were unsuccessful in not only failing to find the exact amount being paid to win a franchise but they also failed to obtain an order of magnitude for these costs. This is because the cost of the franchises is only known to a closed circle of people who are directly involved in the bidding process. This circle consists of people from the train operators who submit the bids for the franchise, the infrastructure manager and the DfT. Furthermore, once a train operator wins a franchise, it receives certain subsidies to provide further services on the network for which the franchise has been awarded and the amount paid as subsidies is also very difficult to obtain. What this means is that if anyone outside the aforementioned circle of people wants to estimate how much it costs to run a service on a network, he will be unable to not only calculate rough estimates for such costs but will also be unable to calculate the order of magnitude of the cost.

Unlike the cost of the franchises, the energy consumption of a train is more readily found and can be estimated with relative accuracy if information is provided about a train's dynamic characteristics (e.g. acceleration and its aerodynamics) as well as terrain characteristics (e.g. gradient) [6]. Such information though is not used when timetabling and, as such, energy minimisation is more accurately calculated using real-time models.

Chapter 4

Optimisation of a railway timetable

4.1 Introduction

An optimisation procedure has been developed to enable the analysis of the cost functions formulated in Chapter 3. The reason for developing a new algorithm rather than relying on one of the multiple existing algorithms is because it is felt that none of the current algorithms can capture the tasks required to carry out the analysis. Existing algorithms start with an already constructed timetable which may be infeasible (due to the occurrence of delays) and carry out rescheduling in order to make it feasible or further improve its quality given the set of objective functions. For the purpose of this project, not only a timetable needs to be

constructed from scratch but, as will be shown in subsequent sections, the number of trains to be scheduled must also vary. It is therefore felt necessary to develop a dedicated algorithm which will allow for the analysis of the cost functions to take place.

The optimisation algorithm developed, evaluates different realisations of a timetable and outputs the one with the lower cost. The results from the algorithm are then validated by entering the output in a simulation environment which is designed to model the movement of trains along the East Coast Main Line.

Section 4.2 describes the optimisation procedure and Section 4.3 the methodology for calculating the passengers on the trains at any point in time. Section 4.4 describes how delays are inserted into the timetable. Section 4.5 explains how the model was validated by both validating the timetable construction method (Section 4.5.4) and the optimisation procedure (Section 4.5.5). Section 4.6 concludes the chapter.

4.2 Description of the optimisation algorithm

The cost functions developed in the previous section are applied to formulate a multi-objective optimisation problem. The optimisation aims to determine the train timetable, in terms of arrival $\tau_{n,i}$ and departure times $\sigma_{n,i}$ for all trains n

over all stations i , such that the following linear combination of costs is minimised:

$$\min_{\tau, \sigma} : C = C_T + C_W + C_P + C_D \quad (4.1)$$

The cost in Equation 4.1 is in monetary units and its cost components are integrated through the monetary cost coefficients c_T , c_W , c_P and c_D we described. The cost minimisation problem is subject to the operational constraints (3.3), (3.4), (3.7) and (3.8).

The train timetable optimisation problem is a combinatorial optimisation problem that involves different feasible combinations of $\tau_{n,i}$ and $\sigma_{n,i}$ representing different sequencing and scheduling of trains [27, 94]. Considering a scenario where there are N trains to schedule, the number of possible sequences for scheduling these trains will be $N!$. This has not included the numerous ways of setting the departure and arrival times of these trains along the service route given a sequence.

To derive a solution within a reasonable time, an optimisation algorithm was developed in Visual C# which works in the stages shown in Figure 4.1. In the first stage a Genetic Algorithm produces a train sequence which is then passed to the second stage which utilises Dijkstra's Algorithm to determine the path of the train through the network. Finally, a Hill-Climbing Algorithm schedules additional trains until the timetable's time span exceeds a predefined threshold. After Dijkstra's Algorithm terminates, the Hill-Climbing heuristic schedules one additional train in each direction and their departure time from their respective

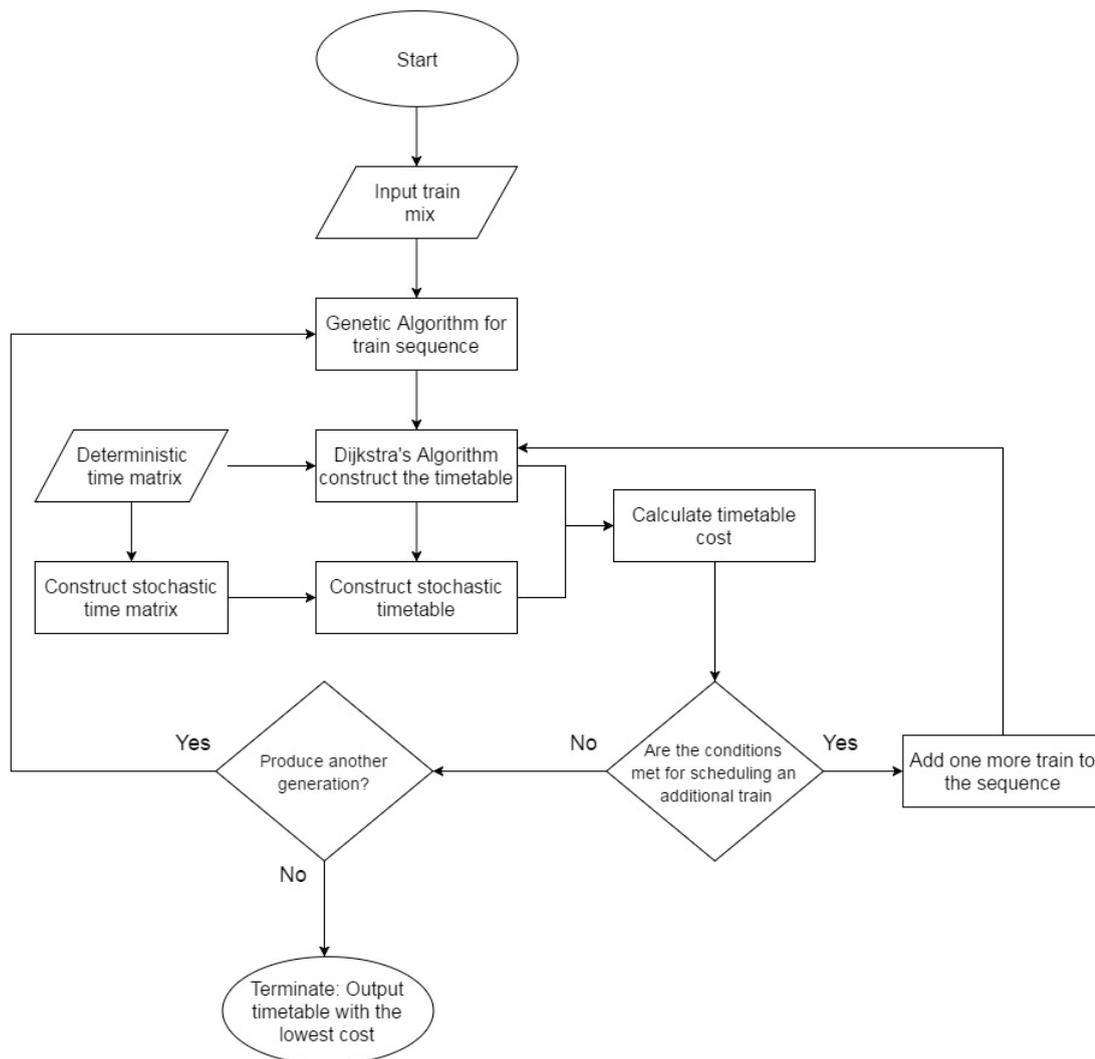


FIGURE 4.1: Flowchart of the optimisation algorithm

origin comes after the departure time of all the previous trains scheduled up to that point. Dijkstra's Algorithm is then re-run to determine arrival and departure times only for the newly added trains and those arrival and departure times are subject to the constraints imposed by all the trains which have been scheduled before the newly added trains. The constraints imposed are the ones defined in Equations (3.3), (3.4), (3.7) and (3.8).

The way the optimisation procedure has been designed means that the optimality of the timetable cannot be guaranteed since neither the Genetic Algorithm nor the Hill-Climbing heuristic are algorithms which are guaranteed to find the optimal solution; they are both approximate methods which may (or may not) return the optimal solution. Nonetheless, the three stage optimisation procedure was designed such that it enables the examination of two of the timetabling characteristics that we want to analyse: train sequencing (controlled by the Genetic Algorithm) and the number of trains on the track (controlled by the Hill-Climbing heuristic). The fact that train sequencing and the number of trains is controlled by different algorithms also allows for the analysis of how the cost of the timetable changes by only changing the sequence or the number of trains while keeping the other constant. Dijkstra's Algorithm is only needed to assign arrival and departure times to the trains.

These stages are further elaborated upon in Sections 4.2.1, 4.2.2 and 4.2.3

4.2.1 First stage - Genetic Algorithm

Genetic Algorithms are based on the concept of natural selection and their use mainly revolves around tackling combinatorial problems for which no efficient algorithms exist [76].

Genetic Algorithms work by encoding possible solutions to the problem as a binary string called *chromosome* while the entries in the binary string are termed

Algorithm 1 Genetic Algorithm pseudo-code

```

1: procedure GENETICALGORITHM(N)
2:   Initialise
3:   Evaluate starting chromosomes
4:   while Termination condition is FALSE do
5:     Select parents
6:     Crossover Parents
7:     Mutate offspring
8:     Create new population
9:   end while
10:  Return best individual
11: end procedure

```

the chromosome's *genes*. The algorithm starts by generating an N number of chromosomes which are then recombined through the process of *crossover*. Crossover is carried out by selecting the chromosomes to be recombined (called *parents*) and then replacing the genes of one parent by the genes of the other in order to generate a new chromosome called *offspring*. For example, assume that we have two parents given in Table 4.1.

TABLE 4.1: Parent chromosomes for crossover

P1	1	0	0	1	0	0	1
P2	0	1	0	0	1	1	0

Assuming a crossover point 4, the offspring is given as the pair in Table 4.2.

TABLE 4.2: Offsprings after crossover

P1	1	0	0	1	1	1	0
P2	0	1	0	0	0	0	1

Following the offsprings' formation, random *mutations* are then inserted, usually by making each gene in the offsprings having a small probability of changing from

1 to 0 or vice versa. Each chromosome in the population is then evaluated using an objective function which assigns a *fitness value* f to each chromosome. The process of elimination then follows which removes the chromosomes with a low fitness value. The chromosomes that survive form the new *generation*. The process of creating new generations continues until a user-defined number of generations is reached upon which the algorithm terminates and returns the chromosome with the higher fitness value [76]. The above procedure is summarised in Figure 4.2 while further details regarding Genetic Algorithms can be found in a number of books and papers including [50, 76, 79].

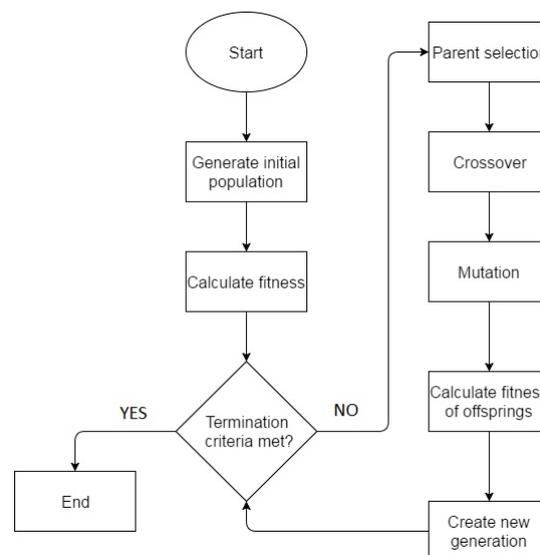


FIGURE 4.2: Genetic algorithm flowchart

Train timetabling involves different feasible combinations representing different sequencing and scheduling of trains at a network's nodes [22, 26, 94]. To derive a solution within a reasonable time, an optimised sequence of trains is searched using a Genetic Algorithm. The Genetic Algorithm starts by generating an initial (random) set of chromosomes with each chromosome representing a different

sequence with which the trains are to be dispatched from their origin. Unlike the traditional binary encoding approach, a permutation encoding scheme is adopted in which each gene within the chromosome represents a train to be scheduled. For example, consider eight trains (A, B, . . . , H) to be scheduled, a total of $8! = 40320$ possible sequences can arise. Each of these 40320 possible combinations can be represented by an 8-bit chromosome. Two possible chromosomes are given in Table 4.3.

TABLE 4.3: Permutation sequencing

A	B	C	D	E	F	G	H
B	A	C	D	E	F	G	H

For the purpose of the project, the initial population is comprised of 200 train sequences. Given that a sequence is produced, the arrival and departure times of the trains is determined using an implementation of Dijkstra’s Algorithm which is further elaborated upon in Section 4.2.2. Evaluating the population’s fitness is carried out using the cost functions described in Chapter 3. Essentially a higher fitness value will be assigned to a train sequence if the resulting timetable achieves lower total cost, and the fitness function FIT_i for each sequence i is defined as:

$$FIT_i = 1 - \frac{C_g}{C_{max}}, \quad (4.2)$$

where C_g is the total cost of a given timetable g and C_{max} is the cost of the most expensive timetable as of the current iteration.

For the reproduction step, the number of train sequences to be selected for crossover is determined using a crossover proportion which is set to 80% meaning that 160 pairs of train sequences are selected for crossover. The parents are then selected using a roulette wheel selection method also known as fitness proportional selection. This method uses a probability distribution for selecting chromosomes based on their respective fit. Random numbers are then used to choose the parents [76]. For example, consider the case of three chromosomes with fitness values 0.7, 0.5 and 0.1 respectively (i.e. each chromosome occupies a section of the roulette proportionate to its fitness value). The roulette can be imagined as being divided into three parts with the first chromosome occupying 54% of the roulette, and chromosome two and three occupying 38% and 8% respectively. A random number x between zero and one is then generated which determines the train sequence to be selected based on the following:

$$\left\{ \begin{array}{l} \text{Train sequence 1} \quad \text{if } x \leq 0.54 \\ \text{Train sequence 2} \quad \text{if } 0.54 < x \leq 0.92 \\ \text{Train sequence 3} \quad \text{if } 0.92 < x \end{array} \right.$$

Furthermore, selection with replacement takes place which means that sequences that lead to the construction of low-cost timetables have a chance of being selected multiple times, increasing the likelihood of generating strong offsprings. The Genetic Algorithm crosses chromosomes over by separating each parent chromosome into two parts, swaps with each other, and forms the new pair of chromosomes.

The mutation process then follows whereby it randomly selects some bits in the

population of train sequences with a predefined probability (in this case a 2% probability is used) and swaps them with another gene within the sequence. This is done to prevent the optimisation process from getting trapped in a local optima.

Once the mutations have been finalised, the fitness of the offsprings is calculated and the process of elimination begins. This process has the task of deleting train sequences which lead to the construction of high-cost timetables to prevent them from crossing over with other sequences. The process of elimination is prohibited from deleting the top 5% of the chromosomes from the previous generation in order to make sure that the current generation is at least as strong as the previous one.

The optimisation process described above (reproduction-crossover-mutation) will continue until the predefined maximum number of iterations (400 generations) is reached. Section 4.5.2 provides the evidence on why 400 generations were chosen as the stopping criterion.

4.2.2 Second stage - Dijkstra's Algorithm

Cormen et al. [24] describe Dijkstra's algorithm as one designed to solve single-source shortest-path problems on weighted, directed graphs. The optimisation process is described below.

Dijkstra's algorithm is a form of Greedy Heuristic but, unlike greedy heuristics

Algorithm 2 Dijkstra's Algorithm pseudo-code

```

1: procedure DIJKSTRASALGORITHM( $G, V$ )
2:   Initialise
3:   Distance from sourceNode to all nodes =  $\infty$ 
4:   Distance from sourceNode to sourceNode = 0
5:   Add all  $v \in G$  to priority queue  $Q$ 
6:   while  $Q$  is non-empty do
7:      $u = Q.removeMin$ 
8:     for all neighbours  $n$  of  $u$  in  $Q$  do
9:       if  $D[u] + w(u, z) < D[z]$  then
10:         $D[z] = D[u] + w(u, z)$ 
11:        Change key of  $z$  in  $Q$  to  $D[z]$ 
12:       end if
13:     end for
14:   end while
15:   Return shortest path
16: end procedure

```

which tend to perform badly when the problem increases in size, Dijkstra's algorithm always return the shortest path on a graph [24].

This stage in the optimisation algorithm determines the $\sigma_{n,s}$ and $\tau_{n,s}$ as the earliest time that each train can travel from origin to destination while considering constraints (3.3), (3.4), (3.7) and (3.8).

Figure 4.3 presents a small network with four nodes and four edges the weight of which indicates the time needed to travel from one node to the next and the headway is 30 units. Assume one train departs from node A at 08:00 and its destination is node D. It is easy to see that the shortest path is via node B in which it is expected to arrive at 08:15 and its arrival time at D is 08:35. Now assume that after the first train, a second train departs from node A at 08:05 and its destination is once again node D. The headway on the edge from A to B means that the earliest the second train can arrive at B is 08:35 so the shortest path from

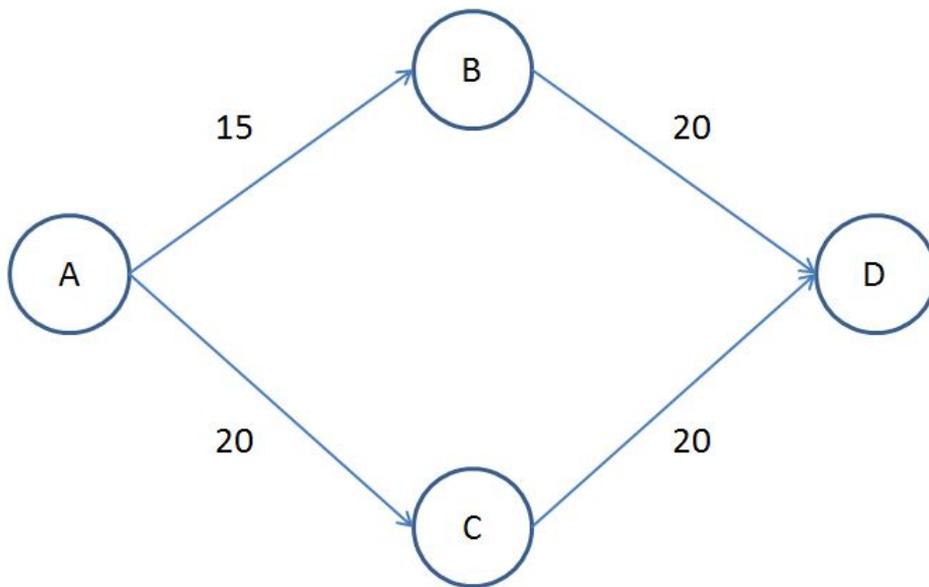


FIGURE 4.3: Sample network

A to D is via C . The implementation in this project incorporates these changes in the distance matrix so all that has to be done is for the algorithm is to determine the shortest path given the time matrix and then update it as necessary given the constrains (3.3), (3.4), (3.7) and (3.8).

Following the calculation of $\sigma_{n,i}$ and $\tau_{n,i}$, the timetable's total cost is calculated as the sum of the timetable's cost of journey time, waiting time, punctuality and crowdedness as defined by the equations 3.17 which are then added as shown in Equation 4.1. Once the timetable's total cost is found, Hill-Climbing is run to determine whether the cost of the timetable can be reduced by scheduling additional trains. This is further explained in Section 4.2.3.

4.2.3 Third stage - Hill-Climbing Algorithm

Hill-climbing algorithms are best described by Russell and Norvig [79] as ‘... a loop that continually moves in the direction of increasing value ... [and] terminates when it reaches a peak where no neighbour has a higher value’.

As mentioned in Section 4.2.2, once the timetable’s total cost with N trains is found, we seek to decrease the cost of the timetable by adding more trains to the schedule. This is done by arranging for two more trains to be added to the schedule; one in each direction (one train in the ‘up’ direction and one in the ‘down’ direction). Arrival and departure times of the two newly added trains is determined by keeping the existing timetable the same and finding arrival and departure times only for the two newly added trains. Consequently the arrival and departure times for the added trains is determined subject to the arrival and departure times of all trains scheduled before them.

For example, assume three trains A, B and C are scheduled which are sequenced by the Genetic Algorithm as

$$B, C, A \tag{4.3}$$

The Hill-climbing heuristic will add a new train D at the end of the above sequence, making the new sequence

$$B, C, A, D \tag{4.4}$$

Arrival and departure times for train D will consequently be calculated using the constraints imposed by the arrival and departure times of trains B, C and A . Now suppose that all trains depart from the same origin in 10 minute intervals and trains B, C and A have been scheduled to depart at 09:00, 09:10 and 09:20 respectively. If a new train D is added to the timetable using the Hill-Climbing heuristic, its departure from its origin will be 09:30 while the departure times for trains B, C and A will remain unaffected.

The rationale for the introduction of this step is that increasing the number of trains to be scheduled will reduce crowdedness but will have an adverse impact on overall punctuality. The additional train is scheduled after all the previous have been scheduled first. There are two termination criteria for the Hill-climb process

- If the timetable's timespan ($\sigma_{1,S_1(1)} - \tau_{N,S_N(I)}$) is exceeded, then the timetable is rendered infeasible
- If the cost of the timetable with N trains is higher than the cost of the previous timetable with $N - 1$ trains

In case that either of the above conditions are met, Hill-climb terminates and returns the cheapest timetable.

4.3 Passenger calculations

As the formulation of the cost functions suggests, information regarding the arrival rate of customers is needed for each origin-destination pair. However, such information is collected by the train operators but is confidential due to data privacy issues and as such it is not available for the public. Therefore, a methodology was developed which allows for the estimation of the number of passengers with relative accuracy.

Section 4.3.1 explains how demand information is summarised in a matrix form and Section 4.3.2 explains how the matrix is used to generate information about the number of passengers on board each train.

4.3.1 Origin-destination matrix

The relative importance of each origin-destination in the network was calculated. This was based on field observations and consultation with the railway industry, allowing for the estimation of approximate values for the proportion of passengers arriving at each station and their destination. Then, based on reports published by Network Rail (e.g. [30]), the average train loading for peak and off-peak hours

was found for a given route. Therefore, by combining the average train loading information with the matrix providing the relative importance of each station, not only the number of passengers on board can be estimated but also their origin/destination can be found in order to calculate the cost functions.

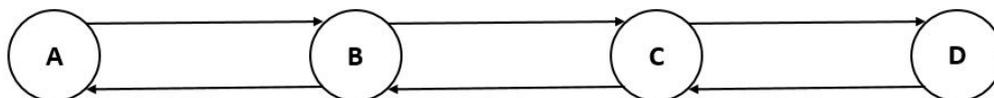


FIGURE 4.4: Sample network for origin-destination matrix illustration

TABLE 4.4: Origin-Destination matrix example

		Destination			
		Station A	Station B	Station C	Station D
Origin	Station A	0	0.2	0.3	0.5
	Station B	1	0	0.4	0.6
	Station C	0.7	0.3	0	1
	Station D	0.5	0.3	0.2	0

Table 4.4 demonstrates an example of how passenger information data will be incorporated in the model based on the network in Figure 4.4. Each element (i, j) in the matrix shows, as a proportion, the number of passengers to board a train at station i going to station j . For example, during morning peaks, trains may be operating at 120% of their seating capacity. This means that if a train with 100 seats arrives at Station B going to Station D with 75 passengers on board, 45 more passengers will board. Table 4.5 shows how those 45 passengers will be divided according to their destination.

TABLE 4.5: Passengers boarded example

Origin \rightarrow destination	Number of passengers boarded
Station B \rightarrow Station C	$45 * 0.4 = 18$
Station B \rightarrow Station D	$45 * 0.6 = 27$

4.3.2 Derivation of train demand

The Origin-Destination Matrix described in Section 4.3.1 is used to calculate the total demand which will be generated during the study period. The algorithm for calculating the total demand generated in the network is given in the pseudo-code below.

Algorithm 3 Total demand calculation pseudo-code

```

1: procedure DEMANDCALCULATION( $N, S_n$ )
2:   for all  $n \in N$  do
3:     for all  $i \in S_n$  do
4:       for all  $j > i$  do
5:         totalDemand( $i, j$ ) += toBoard * ODMatrix( $i, j$ )
6:       end for
7:     end for
8:   end for
9: end procedure

```

The procedure described above derives the matrix with the total demand generated by all trains for all origin-destination combinations. For example, assume that only five trains are considered which travel from Station A to Station D (and vice versa) visiting all stations in between. Each train has a seating capacity of 100 passengers and the average loading factor for trains is assumed to be 100%. The resulting matrix for the total demand is given in Table 4.6.

TABLE 4.6: Demand matrix example for four stations

		Destination			
		Station A	Station B	Station C	Station D
Origin	Station A	0	100	150	250
	Station B	160	0	40	60
	Station C	70	30	0	190
	Station D	250	150	100	0

In order to better understand how Table 4.6 was calculated, one should refer back to Table 4.4. The fact that five trains will arrive at Station A going to Station D and each train has a seating capacity of 100 passengers means that a total of 500 passengers will demand a service from Station A to any subsequent station. The destination of those 500 passengers is determined by referring to Table 4.4. Consequently, if 500 is multiplied by each of the entries in the first row of Table 4.4, the entries in the first row of Table 4.6 are obtained. At Station B, a total of 100 passengers will alight from all trains which leads to 100 seats being vacated which are then filled by the same number of passengers. Using the entries in the right hand side of the diagonal of Table 4.4 gives the values found in the corresponding entries in Table 4.6.

Once the timetable is derived, passengers are allocated to each train scheduled in order to calculate the cost functions. The formula for allocating passengers to train n at each of the stations in its path is given as:

$$p_{n,i}^y = \sum_{\forall j>i} \frac{a_n}{A_{i \rightarrow j}} * totalDemand(i, j) * y\% \quad (4.5)$$

Equation 4.5 shows that the number of passengers to board train n in its journey from station i to station j is a function of the train's seats relative to the total number of passenger seats offered by all trains from station i to station j . Considering the example above where five trains travel from Station A to Station B (and vice versa), the total number of seats for each origin-destination is given by Figure 4.7.

TABLE 4.7: Total seats offered example

		Destination			
		Station A	Station B	Station C	Station D
Origin	Station A	0	500	500	500
	Station B	500	0	500	500
	Station C	500	500	0	500
	Station D	500	500	500	0

The term $y\%$ in Equation 4.5 denotes what percentage of the passengers to board are of type y . When the number of passengers of each type are determined, the final step is to determine which passengers will take a seat. Obviously, this issue only arises when the number of passengers to board exceeds the train's seating capacity. Consequently, when passengers board a train, business passengers are first allocated a seat and if any seats remain, these will be given to commuting passengers. Finally, any remaining seats (if any) are given to passengers who travel for leisure purposes. This is supported by the fact that business and commuting passengers are more likely to plan their trip in advance, making them more likely to reserve a seat for their journey. This assumption is also supported by Network Rail which states that business and commuting passengers tend to plan their trip

in advance [55]. This method for allocating seats to the passengers is much more important than expected since, as Tables 3.3 and 3.4 suggest, the penalties for standing passengers can differ greatly depending on the passenger type. This makes the solutions to the problem very sensitive to the passenger mix, an issue which emphasises the importance of determining the appropriate passenger mix to enter in the model given the hour of the day.

It is important to note that the demand matrix is derived before any additional trains are scheduled through the Hill-Climbing heuristic. This achieves the purpose of scheduling additional train without increasing demand at the same time, enabling for the examination of the reductions in the levels of crowdedness.

4.4 Punctuality modelling

In Chapter 2, three different methods have been identified which model the uncertainty in railway timetables: queueing models, analytical models and simulations. For the purposes of this experiment, it has been decided to model the randomness in the running times of the trains using simulation. Queueing models were ruled out since they are timetable independent, rendering them inappropriate for timetable evaluation purposes. Furthermore, the computational intractability of many analytical models (e.g. [16, 96]) leads to the use of heuristics approximations which is likely to reduce the quality of the outcome.

The procedure to incorporate delays into the timetable has been developed as follows. Primary delays will be modelled using an exponential distribution, the Probability Density Function of which is given as:

$$\mu e^{-\mu x}, \quad (4.6)$$

Where x represents the sectional running time over any given section and μ is the rate parameter. When the optimisation procedure (Figure 4.1) is initialised, and before any Genetic Algorithm generations are created, the stochastic time matrices are calculated as shown in Figure 4.5.

The process starts by filling the two-dimensional arrays which contain information about the deterministic running times of the trains along all edges in the network. The running times will differ depending on the train class type (e.g. class 75, class 442) so a time matrix exists for each train class.

To better understand how a time matrix is constructed, consider the small network in Figure 4.6 which consists of a single track going from Node 1 to Node 2 via the signalling blocks b and c . The time needed to traverse the distance from block b to block c is 5 time units and from signalling block c to Node 2 is 7 time units. The time-space diagram for this movement is shown in Figure 4.7. The time matrix constructed from the information provided in Figure 4.6 is shown in Table 4.8.

The two-dimensional time matrix array will have three entries in each dimension and the non-zero entries will correspond to the feasible links in the network. The

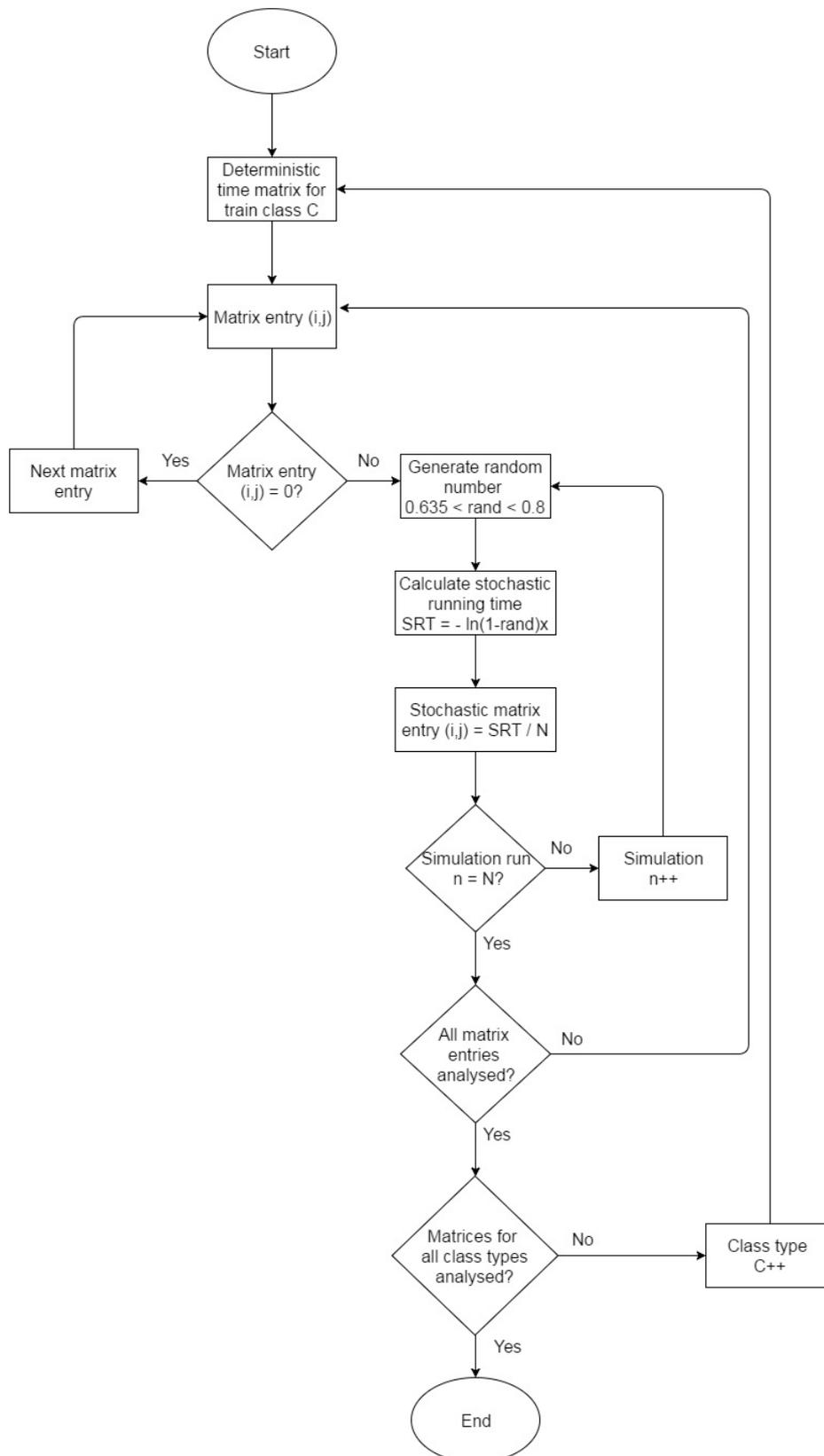


FIGURE 4.5: Construction of the stochastic matrices flowchart

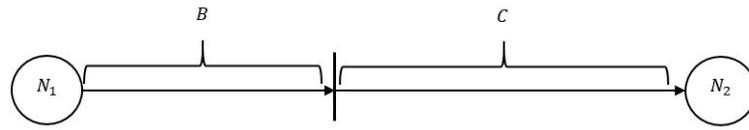


FIGURE 4.6: Network for time matrix illustration

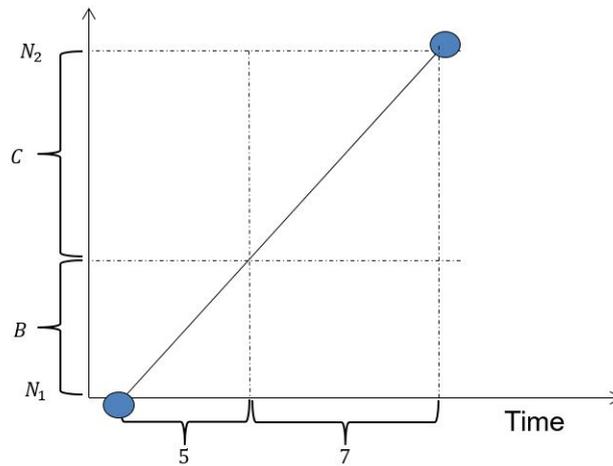


FIGURE 4.7: Time-space diagram for time matrix illustration

TABLE 4.8: Time matrix example

	B	C	N_2
B	0	5	0
C	0	0	7
N_2	0	0	0

stochastic running time matrix of each class type is then derived from the deterministic time matrices. This is done by examining each entry in the deterministic matrix and if the entry is zero, no feasible link exists and the algorithm moves on to examine the next entry in the matrix. If the entry is non-zero, a feasible edge exists and the stochastic running time for that edge is calculated. The stochastic

running time is calculated by generating another random number in the range

$$0.635 \leq rand < 0.8 \quad (4.7)$$

which is then used to calculate the stochastic running times $S\tilde{R}T$ using the equation

$$S\tilde{R}T = -\ln(1 - rand) * \mu, \quad 0.635 \leq rand < 0.8 \quad (4.8)$$

The above procedure is repeated an N number of times (in this project $N = 100$) and the average value of the simulation runs is then entered in the stochastic time matrix. The λ parameter of the exponential distribution ($\lambda = \mu^{-1}$) and the range of values of the $rand$ variable were chosen such that the base case (i.e. where no extra trains are added) stochastic timetable meets the punctuality metrics published by Network Rail [60]. This procedure is repeated until a stochastic running time is computed for each non-zero entry in the time matrices of each train class. After this process terminates, each train class has two time matrices: one deterministic and one stochastic.

The above procedure ensures that the same stochastic matrices are used in all the generations produced by the Genetic Algorithm. This ensures that the Genetic Algorithm will converge.

It is understood that the large number of times that Equation 4.8 is run (i.e. 100

times) and then averaged means that the elements to be entered in the stochastic matrix will be close to their average values (following from the law of large numbers); this, in turn, has one important implication. If two stochastic time matrices are generated and an arbitrary element ($\eta_{i,j}$) is selected from matrix η and compared to the corresponding element ($\theta_{i,j}$) from matrix θ , it will be observed that the two elements will be very close (i.e. $\eta_{i,j} \approx \theta_{i,j}$). This is because, as mentioned above, the large number of iterations will fill the matrices with values which are close to the average sectional running times. Although this suggests that the timetable generated does not exhibit much variability, this is actually desirable since the purpose of this procedure is to generate small disturbances rather than large scale disruptions. This decision can be justified since timetables are not designed to cope with big disruptions for two reasons. The first is that the magnitude of the disruption is unknown making it difficult to compensate for it a-priori and the second is that if the timetable is designed such that it minimises the impact of large scale disruptions, robust optimisation methods will be used which, as shown in Section 2.2.2.3, provide undesirably conservative solutions. Therefore, timetables are usually designed such that they absorb small disturbances while large scale disturbances are dealt with real-time using specialised algorithms (e.g. [46, 49]). Therefore, multiple runs of Equation 4.8 had to be taken to calculate the average in order to prevent the scenario where an extremely large delay was generated from the exponential distribution. Nonetheless, if the need ever arises to make the timetable truly stochastic by inserting more volatile sectional running times, the number of times that Equation 4.8 is run and then averaged can be

reduced. For example, if Equation 4.8 is only run once and the running times generated are used to construct the stochastic running times matrix, the timetable which will be created will exhibit much higher variability.

After the stochastic time matrix for all train types is constructed, the optimisation algorithm described above is initiated. When Dijkstra's Algorithm constructs a deterministic timetable, the timetable is recalculated using the stochastic time matrix. Trains are dispatched using the same sequence as before, and the path they follow is exactly the same as the one determined by Dijkstra's Algorithm. In one of the locations in their path the trains experience delays and the delay is incorporated by referring to the time in the stochastic time matrix. The location that each train is delayed is chosen by generating a random number which refers to one of the signalling blocks in each train's path. When this procedure is repeated for all trains, the algorithm will output two timetables: one deterministic and one stochastic. The time deviation of a train's stochastic timetable from the deterministic timetable is the amount of time that the train is delayed.

It should be stressed that this procedure does not change the initial train order in any way, it only introduces small amount of noise to the deterministic timetable. It is also important to understand that the scope is to only consider very small small disturbances since no timetable can be proactively prepared to recover from large scale disruptions since such disruptions are dealt with real time.

4.5 Model validation

Model validation is important in order to make sure that the solutions the optimisation algorithm generates are feasible in a real life context and, if deemed necessary, fine-tune the model if deemed necessary. This is achieved by generating timetables for a section on the East Coast Main Line (ECML) which is then input into the BRAVE simulation. Section 4.5.1 provides an outline for the network to be used for validation and Sections 4.5.3 and 4.5.4 describe the BRaVE simulation environment and the results of the validation process respectively. Lastly, Section 4.5.5 validates the convergence of the optimisation procedure.

4.5.1 East Coast Main Line

The ECML is a part of Network Rail's Route G and provides the most direct, high-speed connection between London and Edinburgh [62]. The main part of the line is being powered through overhead electrification [62].

The route serves several high-speed intercity services such as from London to Leeds. On top of intercity services, the ECML provides a number of important local services such as the Moorgate Branch and the Hertford loop which experience heavy congestion especially during peak hours [62, 63]. The ECML is an important route for freight trains which mainly operate on the northern part of the route,

parallel to the A1 motorway. In the southern part of the route freight trains often utilise the Hertford loop [62].

Train operators operating on the ECML network include among others Virgin Trains East Coast, East Midlands Trains and CrossCountry. Due to the enormous size of the ECML and the numerous operators utilising it, a wide range of both passenger and freight rolling stocks can be found. For example, local services are usually run by 313, 317 and 321 classes while intercity services use the class 125 High Speed rolling stock [67].

4.5.1.1 Alexandra Palace to Hatfield section

Modelling a big network such as the ECML is considered impractical due to the time needed to prepare the optimisation model as well as due to the huge increase in computational time. Therefore, a smaller section of the network has been chosen which is illustrated in Figure 4.8 enclosed in a green square. The section includes all the stations between Alexandra Palace and Hatfield but in the Hertford loop, only Bowes Park is considered. This section was chosen due to the fact that it has the necessary complexity both in terms of multiple train paths as well as the heterogeneity of train services operating on it.

Between Alexandra Palace and Hatfield four tracks exist, two either way. The outer two tracks in each direction are used by local services while intercity services run non-stop on the inner tracks. Between Alexandra Palace and Bowes Park, one

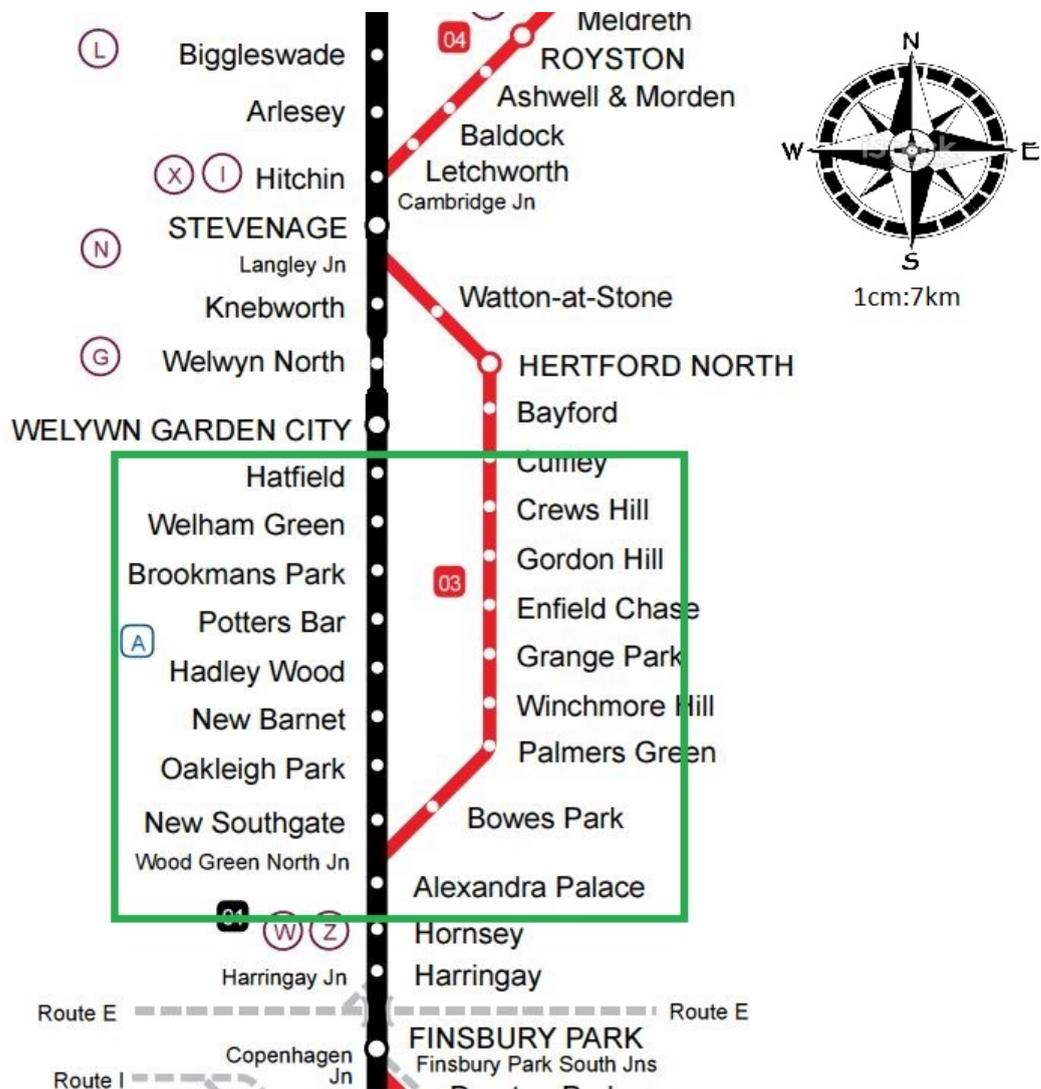


FIGURE 4.8: ECML between Alexandra Palace and Hatfield

line exists in the 'down' direction (towards Welwyn Garden city) and two lines in the 'up' direction (towards London) with one of the lines passing through the Bounds Green depot ¹.

For the purposes of the simulation, the train mix between the weekday hours of 08:00-09:00 will be used. Information about the train mix was collected from Network Rail's working timetable [67] and the results are summarised in Table

¹The railway operations in the United Kingdom refer to routes being in the 'up' direction if they lead towards London and in the 'down' direction if they lead away from London

4.9. No freight trains operate on the network during the 08:00-09:00 time interval potentially due to the need to minimise the risk of disruptions caused by the increased speed heterogeneity during the morning peak.

TABLE 4.9: Train mix between 08:00-09:00 on a weekday

	Class 313		Class 317		Class 125	
	Down	Up	Down	Up	Down	Up
Alexandra Palace - Hatfield	8	5	10	8	8	7
Alexandra Palace - Bowes Park	8	6	0	0	0	0

Local services are run by class types 313 and 317 while intercity services are run by class type 125. It should be noted that, for simplification purposes, only 313 class types stop at all stations while all semi-fast services utilise a 317 rolling stock (Table 4.9). This has very little impact on the quality of the experiments but it considerably speeds up the optimisation algorithm.

Data regarding the sectional running times of all class types was collected by referring to the working timetable [67] and, if the need arose, fine-tuned further during the validation process.

Finally, information needed to fill the origin-destination matrix was gathered through field observations, consultation with industry professionals and passenger statistics published by the ORR [61]. Appendix A shows the matrix constructed from the information collected.

4.5.2 Algorithm convergence - ECML network

An initial analysis was run on the network to determine whether the algorithm converges and how quickly it does so. This was done in order to fine-tune its parameters such that a good quality solution is achieved within a reasonable amount of time. The analysis was carried out on the East Coast Main Line network for demand levels equal to 100% of the available seats. The train mix used is the one given in Section 4.5.1.1 and the criteria to consider the algorithm as having converged is for five consecutive improvements to improve the cost of the timetable by less than 0.5% or no further improvements take place after 50 iterations.

```
TIIE:///C:/Users/ansp/vropoox/university/FNU/ train sceneu
Generation: 0, Timetable Cost: 125832.9
Generation: 1, Timetable Cost: 113389.5
Generation: 2, Timetable Cost: 113389.5
Generation: 3, Timetable Cost: 113389.5
Generation: 4, Timetable Cost: 112063.2
Generation: 5, Timetable Cost: 109896.6
Generation: 6, Timetable Cost: 109896.6
Generation: 7, Timetable Cost: 109896.6
Generation: 8, Timetable Cost: 109896.6
Generation: 9, Timetable Cost: 109751.2
Generation: 10, Timetable Cost: 109751.2
Generation: 11, Timetable Cost: 107480.4
Generation: 12, Timetable Cost: 107480.4
Generation: 13, Timetable Cost: 107181.7
Generation: 14, Timetable Cost: 107181.7
Generation: 15, Timetable Cost: 107181.7
Generation: 16, Timetable Cost: 106205.1
Generation: 17, Timetable Cost: 106205.1
Generation: 18, Timetable Cost: 106205.1
```

FIGURE 4.9: Initial iterations of the optimisation algorithm

Figure 4.5.2 indicates that 400 iterations provides for a more than an adequate termination criterion for the algorithm. This is because the last iteration which led to an improvement of more than 0.5% was the 149th iteration. Furthermore, the last 200 iterations only had one occurrence where any improvements were achieved. Even though the algorithm seems to perform well for 150 generations,

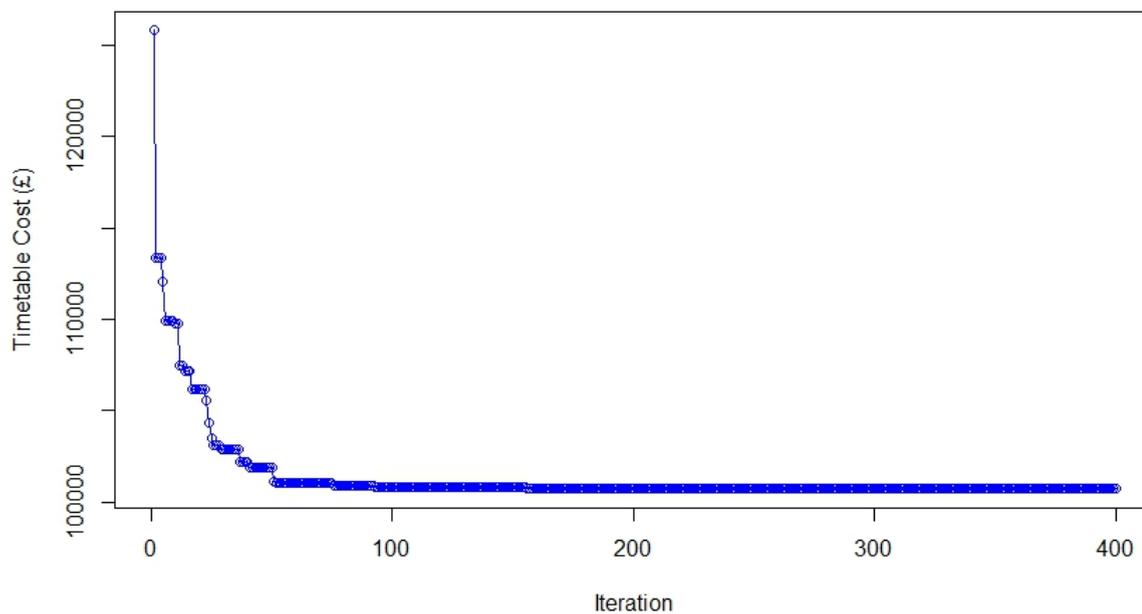


FIGURE 4.10: Algorithm convergence after 400 iterations

such tests will be carried out again in Chapter 5 since both the network and the train mix will change completely.

4.5.3 BRaVE simulation environment

BRaVE is a microscopic simulation environment developed by the University of Birmingham and is capable of simulating all the basic functionality of railway systems [93]. A user is able to define numerous parameters such as infrastructure data, rolling stock characteristics, interlocking arrangements and timetable information [93]. For example, infrastructure data refers to the physical layout of the network, rolling stock characteristics defines the acceleration, deceleration

and top speed of all different train types and the timetable is entered by defining arrival/departure times for all stations in the trains' path.

One of the main reasons for developing BRaVE was to address the need to assess a timetable's performance with respect to the energy consumed and punctuality. Energy consumed is calculated by considering several dynamic characteristics such as acceleration and deceleration rates and terrain characteristics. Punctuality is evaluated by introducing either systematic or random delays. Systematic delays are introduced by selecting the driving profile of each train's driver (e.g. slow, fast). Each profile introduces a systematic variation to the running times of the trains. Random delays are entered by using a seeded random number generator which increases the dwell time of the trains by a random value between 0 – 15 seconds. The seeds can be stored so that further simulations can be carried out by using the same set of random numbers.

A problem with BRaVE is that it is unable to calculate the cost of any of the cost functions the way they have been defined in Chapter 3. Therefore, the option offered by BRaVE to insert delays was not utilised during the validation process. This means that BRaVE was only used as platform to enter the deterministic timetable constructed by the optimisation algorithm described above and check whether the timetable can be replicated in BRaVE. Consequently, BRaVE is used to make sure that the algorithm constructs feasible timetables but a different method needed to be devised to make sure that the timetables the optimisation algorithm constructs are optimised (4.5.5).

4.5.4 Validation of timetable construction method in BRaVE

Model validation is carried out by generating a timetable using the optimisation algorithm described above and entering it in BRaVE. A simulation run is then carried out and an output is produced by BRaVE which provides the arrival and departure time of all trains from the stations. The output generated from BRaVE is not necessarily the same as the timetable entered, the two can differ in cases where the timetable entered in BRaVE is infeasible. If an infeasible timetable is entered, BRaVE has the flexibility of altering the timetable in order to make it abide by the feasibility criteria. Finally, the timetable produced by the optimisation algorithm is compared to that generated by BRaVE and if the two timetables match, the optimisation algorithm is deemed to produce feasible timetables. The validation process is summarised in Figure 4.11.

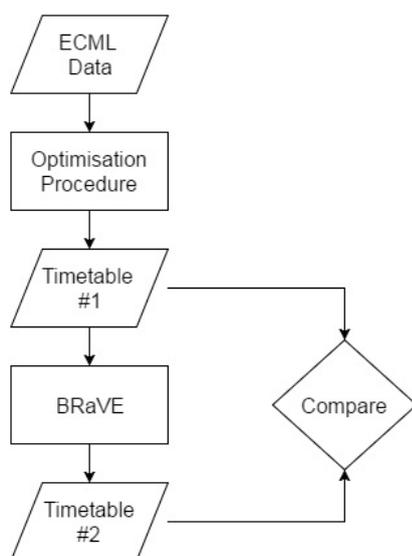


FIGURE 4.11: Validation process flowchart

The validation process starts by feeding the ECML data outlined in Section 4.5.1

into the optimisation algorithm and using it generate timetables. An small excerpt of the timetable is shown in Table 4.10 and a more detailed version in provided in Table B.1 of Appendix B.

TABLE 4.10: Timetable excerpt from the optimisation algorithm

Service	Alex. Palace	Bowes Park	N. Southgate	Oakleigh Park
S62	08:00:00		08:03:00	08:06:30
S857	08:03:00		08:06:00	08:09:30
S31	08:29:05		08:26:05	08:22:35

The timetable is then input in BRaVE which is subsequently run to generate a report with the arrival times as calculated by BRaVE. A small sample from the report is summarised in Table 4.11 and a more detailed version is given in Table B.2 of Appendix B.

TABLE 4.11: Timetable excerpt from BRaVE

Service	Alex. Palace	Bowes Park	N. Southgate	Oakleigh Park
S62	08:00:00		08:03:00	08:06:37
S857	08:03:00		08:05:51	08:09:18
S31	08:28:54		08:25:54	08:22:28

Tables 4.10 and 4.11 show that the two timetables have slight differences with respect to the arrival times of trains at the stations. Initially, the timetable produced from the optimisation algorithm had larger deviations from BRaVE's timetable and a closer examination showed that this was caused by significant differences in the running time of the trains between stations. This was caused by the fact that the time matrix used by the optimisation algorithm to calculate the values for $\tau_{n,i}$

and $\sigma_{n,i \rightarrow j}$ were significantly different from the actual time needed by trains to traverse the same distance as calculated by BRaVE. Another source of variation was the fact that BRaVE records arrivals when a train stops at a station while the optimisation algorithm records arrivals when a train starts occupying a signalling block. This means that the optimisation algorithm and BraVE have inherently different methodologies for recording delays, meaning that an exact match between the two models is impossible to be achieved. A few simulation runs were carried out in BRaVE in order to collect information to enable the construction of a more representative time matrix to be used by the optimisation algorithm. As a consequence, the running times in BRaVE were observed in more detail which led to a update of the data used by the optimisation algorithm such that they closely match the data from BRaVE. This process was iterated multiple times (Table 4.11) with each iteration providing more accurate sectional running times. This process was terminated when the timetable constructed by the optimisation algorithm was almost the same as the one produced by BRaVE.

As evident from Tables 4.10 and 4.11 small differences in the two timetable persist and this can be attributed to two factors. The first one is the difference in the method of calculating the sectional running times of the trains. The optimisation algorithm calculates the timetable by referring to a time matrix while BRaVE calculates the running times dynamically by utilising information about the acceleration/deceleration rates, top speed and terrain characteristics. It is obvious that BRaVE has a more sophisticated and more accurate method of calculating running times which cannot be replicated by the optimisation algorithm due to

its inability to incorporate the dynamic characteristics of the trains. Efforts were made to construct the time matrix as accurately as possible but small discrepancies are expected to persist due to the different methods employed by the optimisation algorithm and BraVE. The second factor is the difference of the two models in calculating the arrival time at stations which is illustrated in Figure 4.12.

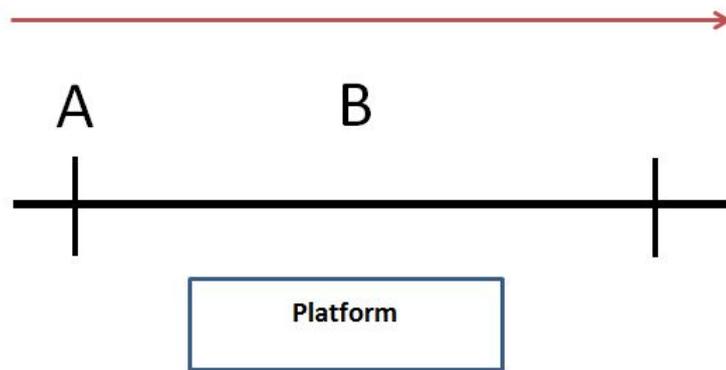


FIGURE 4.12: Monitoring point recording the arrival of trains for the optimisation algorithm (A) and BRaVE (B)

The arrival time at a station as given by the optimisation algorithm is the time the train enters the station tracks (point A) while the arrival time in BRaVE indicates the time when a train stops at the station (point B). The optimisation algorithm reports the arrival time at any node in its path as the time that the node will be marked as occupied by the specific train (Equation 3.5). This was deemed necessary as it allowed the model to quickly identify which nodes were available or occupied at any point to prevent to trains from occupying the same signalling block simultaneously. On the contrary, BRaVE is a much more sophisticated

software which can store information regarding single-block occupancy while also being able to report the actual time that a train stops at a station.

Despite the fact that the optimisation algorithm and BRaVE work in a way that makes it unlikely to provide identical output, the time discrepancies evident in Tables B.1 and B.2 in Appendix B were deemed to be insignificant due to their small magnitude. This implies that the timetables generated by the optimisation algorithm are feasible.

4.5.5 Validation of the optimisation procedure

This chapter focuses on validating the optimisation procedure to make sure that the timetables it constructs are indeed optimised. Unlike Section 4.5.4 which used BRaVE for validation, this section required a different approach due to the fact that BRaVE does not have a method of calculating the cost of the timetable as defined by the cost functions formulated in Chapter 3. The adopted approach aimed to identify whether the timetable sequence provided by the Genetic Algorithm and the optimised number of trains introduced by the Hill-Climbing heuristic do indeed produce timetables of lower cost.

Checking whether the timetable produced by the algorithm was indeed optimised was achieved by finding the optimal solution by manually trying all possible solutions (i.e. brute force) and then comparing it to the result produced by the

optimisation algorithm. Furthermore, in order to make it easier to find the optimal timetable using brute force, only five trains were considered which only travel in the direction from Alexandra Palace to Hatfield (Figure 4.8). All five trains scheduled are class 313 and four out of the five trains travel from Alexandra Palace to Hatfield while stopping at all stations in between while the fifth train travels from Alexandra Palace to Bowes Park. Moreover, the passenger demand was set at 100% of train seats leading to the construction of the matrix in Figure 4.13. Finally, the delayed running times of the trains were calculated once and then used in both brute force calculations and in the calculations run by the algorithm in order to ensure that the results were comparable.

		DESTINATION									
		Alexandra Palace	Bowes Park	New Southgate	Oakleigh Park	New Barnet	Hadley Wood	Potters Bar	Brookmans Park	Welham Green	Hatfield
ORIGIN	Alexandra Palace	0	232	61	61	61	61	93	61	61	371
	Bowes Park	0	0	0	0	0	0	0	0	0	0
	New Southgate	0	0	0	4	4	4	12	4	4	30
	Oakleigh Park	0	0	0	0	3	3	12	3	3	38
	New Barnet	0	0	0	0	0	4	12	4	4	40
	Hadley Wood	0	0	0	0	0	0	7	7	7	51
	Potters Bar	0	0	0	0	0	0	0	14	14	111
	Brookmans Park	0	0	0	0	0	0	0	0	9	84
	Welham Green	0	0	0	0	0	0	0	0	0	104
	Hatfield	0	0	0	0	0	0	0	0	0	0

FIGURE 4.13: Passenger matrix used for algorithm validation

The first section of the algorithm to be validated was the resequencing stage carried out by the Genetic Algorithm. Since four out of the five trains are essentially the same service (i.e. rolling stock 313 from Alexandra Palace to Hatfield) and one train goes to Bowes Park, there were only five different ways the sequence could be set up and these possible sequences are shown in Table 4.12.

TABLE 4.12: Possible train sequences for algorithm validation

Sequence 1	A	B	B	B	B
Sequence 2	B	A	B	B	B
Sequence 3	B	B	A	B	B
Sequence 4	B	B	B	A	B
Sequence 5	B	B	B	B	A

In Table 4.12, *A* refers to the train from Alexandra Palace to Bowes Park while trains which travel from Alexandra Palace to Hatfield are denoted as *B*. The timetables arising from the sequences in Table 4.12 were then found using brute force and their cost calculated. The results from the brute force calculations are shown in Table 4.13.

TABLE 4.13: Timetable cost for all sequences using brute force

		Timetable Cost (£)
Sequence 1	A B B B B	11305
Sequence 2	B A B B B	9603
Sequence 3	B B A B B	8913
Sequence 4	B B B A B	10725
Sequence 5	B B B B A	11305

It is therefore apparent that the sequence

$$BBABB \tag{4.9}$$

is the one which provides the best solution. The next step was to enter the same train mix in the optimisation algorithm and allow it to run in order to see what train sequence it will consider as the one with the best fit. The output from the algorithm agreed with the results from the brute force experiments in that the

sequence in Equation 4.9 leads to the construction of the most efficient timetable at a cost close to the one calculated by the brute force procedure². Nonetheless, slight discrepancies were expected in the calculation of the cost of the timetable since the calculations for the brute force procedure were carried out by hand and involved a great deal of rounding which was the reason for the slight difference in the total cost of the timetable. Consequently, since both the brute force procedure and the optimisation algorithm produced the same sequence, it was concluded that the sequence generated by the Genetic Algorithm does indeed tend to converge to the optimal timetable.

The second section of the algorithm to be validated was the Hill-Climbing heuristic and whether or not it operates in such a way that it converges to the optimal timetable. Similar to the validation of the Genetic Algorithm, the optimal solution for the small instance of the problem was found using a brute force procedure and then compared to the solution given by the optimisation algorithm. Since the instance of the problem used in validation was composed of only five trains, the maximum span of the timetable was set to 45 minutes. This meant that only a handful of additional trains could be scheduled before the span of the timetable exceeded the 45 minute constraint. Furthermore, the trains added were 313 classes which travel from Alexandra Palace to Hatfield while stopping at all intermediate stations. The sequence used to validate the Hill-Climbing heuristic is the one in Equation 4.9 and the results from the brute force procedure are given in Table

²The cost of the timetable calculated by the brute force procedure was £8913 while the cost of the timetable calculated by the optimisation algorithm was £8927

4.14.

TABLE 4.14: Timetable cost for the number of trains scheduled (Brute force)

Number of trains scheduled	Timetable Cost (£)
5 Trains	8913
6 Trains	8472
7 Trains	8057
8 Trains	8092

The solution derived from the brute force procedure shows that scheduling two additional trains lead to the most efficient timetable with a cost of £8057. The train sequence in Equation 4.9 was then inserted in the optimisation algorithm and the Hill-Climbing heuristic was run to schedule additional trains without changing the initial sequence. The results obtained are summarised in Table 4.15.

TABLE 4.15: Timetable cost for the number of trains scheduled (Optimisation algorithm)

Number of trains scheduled	Timetable Cost (£)
5 Trains	8927
6 Trains	8482
7 Trains	8061
8 Trains	8108

Table 4.15 is consistent with Table 4.14 in the sense that when two additional trains are scheduled the cost of the timetable is minimised. The results differ slightly but this can once again be attributed to the effect of rounding. Therefore, both the Genetic Algorithm and the Hill-Climbing heuristic are deemed to produce timetables which converge to optimality.

The above two validation checks ensured that both the Genetic Algorithm and Hill-Climbing heuristic, when run separately, can produce the optimal solution for the problems they examine (i.e. trains sequence and number of trains). Therefore, the final stage of the optimisation algorithm validation was aimed at determining whether the combination of Genetic Algorithm and Hill-Climbing heuristic does indeed return the optimal timetable. Following the same methodology used until this stage, the optimal sequence combined with the optimal number of trains was found first by trying all possible solutions. The results are shown in Table 4.16. The results from the algorithm are summarised in Table 4.17.

TABLE 4.16: Timetable cost per sequence and optimal number of trains (Brute force)

Sequence	Optimal number of trains	Timetable Cost (£)
ABBBB	1	10585
BABBB	2	8920
BBABB	2	8057
BBBAB	2	9288
BBBBA	2	10031

TABLE 4.17: Timetable cost per sequence and optimal number of trains (Optimisation Algorithm)

Sequence	Optimal number of trains	Timetable Cost (£)
ABBBB	1	10601
BABBB	2	8932
BBABB	2	8061
BBBAB	2	9297
BBBBA	2	10040

Both the results from the brute force procedure (Table 4.16) and optimisation algorithm (Table 4.17) agree that the optimal timetable is the one where trains

are sequenced in the order given in Equation 4.9 where two additional trains are added to the sequence meaning that the optimised timetable will look like

$$BBABBBB \quad (4.10)$$

Finally, since in all three validation checks the result from the optimisation procedure matched the results from the brute force procedure, it can be deduced that the optimisation algorithm is valid.

4.6 Summary

An optimisation algorithm is developed which allows for the examination of the cost functions formulated. The optimisation algorithm works in three stages with the first stage being an implementation of a Genetic Algorithm producing a sequence with which trains are to be dispatched from their origin station. Dijkstras Algorithm then constructs the timetable by determining the shortest path between the origin and the destination of each train subject to the constraints of minimum sectional running times, headway and single train occupancy of each block. The third stage of the algorithm schedules additional trains until the timetable either becomes infeasible or the cost of the timetable cannot be minimised any further.

The optimisation model is validated by comparing its output with the timetable produced by the BRaVE simulation software. To enable the comparison, a timetable

is constructed on the ECML network using the train mix traversing the Alexandra Palace to Hatfield subsection between 08:00 to 09:00. A timetable from the optimisation algorithm is then input in BRaVE in order to see if the timetable is feasible. The results from BRaVE show that the result from the optimisation algorithm can be replicated in BRaVE without any significant changes. Small deviations from the timetable entered in BRaVE are observed with a magnitude of a few seconds but these deviations can be attributed to the fact that BRaVE calculates travelling times dynamically while the optimisation model relies on a distance matrix. Consequently, it can be argued that the timetable constructed by the optimisation algorithm is feasible.

Finally, Section 4.5.5 validates the optimisation procedure by taking a small instance of the timetable with five trains and then finding the optimal solution by manually calculating all possible scenarios. Once the optimal solution is found, the algorithm is run on the same instance of the problem and the results it produces are compared to the ones obtained manually. It is shown that the results from the algorithm do indeed match the optimal solution obtained from manual calculations, supporting the argument that the algorithm returns timetables which converge to optimality.

Chapter 5

Case study

5.1 Introduction

The cost functions and the optimisation algorithm were applied on a subsection of the Brighton Main Line in order to tackle the research questions stated in Section 1.3. The information is provided by initially explaining the context of each research question, followed by the graphs summarising experimental results and finally discussing the results obtained.

Section 5.2 describes the network in terms of its physical characteristics, train mix and the passenger demand between each origin-destination pair. Section 5.3 serves the purpose of determining how many generations of the Genetic Algorithm are required before the algorithm converges to a solution. Section 5.4 answers the

question of whether the timetable cost can be reduced by only resequencing the trains. Section 5.5 determines what is the impact on the timetable when additional trains are scheduled. In Section 5.6 the interaction between the cost functions is further analysed by constructing the Pareto Frontier of different pairs of cost functions. The chapter concludes in Section 5.7.

5.2 Brighton Main Line

The network used for the experiments is composed of a subsection of the Brighton Main Line. The Brighton Main Line is approximately 80-km long electrified connection linking London Victoria and London Bridge with Brighton via East Croydon and Gatwick Airport. The line itself has a complex structure with a variable number of tracks (four tracks from London down to Balcombe Tunnel Junction and two tracks thereafter), different speed limits along the line, multiple branch lines (e.g. at Junctions Horsham, Lewes), and sidings (e.g. along Ardingly, Lovers Depot). Govia is the primary passenger operator that operates on the BML. [35].

For practicality purposes, we chose to model only the section between Gatwick Airport and Brighton which is highlighted in Figure 5.1. This is one of the busiest sections along BML and the Keymer Junction is a flat junction identified as one of the network's major bottlenecks [30]. The study period is 08:00 - 10:00, which is regarded as the morning peak, on weekdays. During the study period, a total of 22 trains run from Brighton toward Gatwick and hence Central London (the

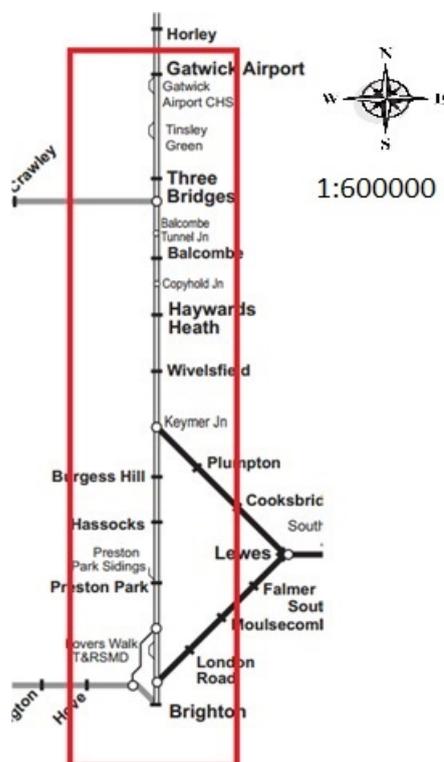


FIGURE 5.1: Brighton Main Line between Gatwick Airport and Brighton

'Up' direction) and 18 trains running from Gatwick toward Brighton (the 'Down' direction) (Table 5.1).

TABLE 5.1: Train mix between 08:00-10:00 on a weekday

	Class 375	Class 442	Total
Gatwick Airport → Brighton	10	8	18
Brighton → Gatwick Airport	14	8	22

The 'base case' train timetable is derived from information obtained from Network Rail. The idea is to derive an optimised timetable from the proposed optimisation framework with the same number of trains within the same study period. We then compare the 'optimised' timetable with this 'base case' timetable to see how much improvement, in terms of reduction in costs, can be achieved in different

aspects through re-sequencing and re-scheduling. There are two different train classes running through the section during the study period: Classes 375 and 442 with Class 375 used for the express connection (Gatwick to/from Brighton with no intermediary stops).

Similar to the East Coast Main Line in Section 4.5.1 information regarding the passengers was collected by referring to reports published by Network Rail and through consultation with industry professionals. Considering the fact that the study will cover the time period 08:00 - 10:00, the origin-destination matrix was constructed (Appendix C) and the passenger mix has been decided to be set as follows:

TABLE 5.2: Passenger mix for the time period 08:00 - 10:00

Passenger Type	
Business	20%
Commuter	60%
Leisure	20%

The origin-destination matrix matrix constructed for the case were the loading factor equal 100% is shown in Figure 5.2.

		DESTINATION								
		Gatwick	Three Bridges	Balcombe	Haywards Heath	Wivelsfield	Burgess Hill	Hassocks	Preston Park	Brighton
ORIGIN	Gatwick	0	3608	230	230	230	230	230	230	3340
	Three Bridges	4157	0	97	97	97	97	97	97	390
	Balcombe	248	61	0	32	32	32	32	32	109
	Haywards Heath	127	79	55	0	46	46	46	46	121
	Wivelsfield	833	505	166	166	0	69	69	69	137
	Burgess Hill	82	50	9	9	9	0	120	120	159
	Hassocks	78	46	9	9	9	9	0	120	318
	Preston Park	779	467	61	61	61	61	61	0	477
	Brighton	3496	936	103	103	103	103	103	103	0

FIGURE 5.2: Origin-destination matrix for a 100% loading factor

Validation for the figures presented in Appendix C, Table 5.2 and Figure 5.2 was carried out both through industry consultation and site visits.

With respect to punctuality, the parameter Φ in Equation 3.13 is set equal to zero. This means that any deviations from the scheduled arrival time will be penalised, irrespective of their magnitude. The decision to set Φ equal to zero was taken after consultation with industry professional and despite the fact that the current industry standards assume all deviations of less than three minutes do not incur any penalties [59, 60, 66]. Section 4.4 explains how the exponential distribution will be used to model delays so, in order to model delays as accurately as possible, the parameter λ in Equation 4.6 was taken to have a value of 1.1 for all the delayed trains. This ensured that delay statistics are in line with the figures published by Network Rail [60].

5.3 Algorithm convergence - BML network

Following the network's incorporation into the optimisation model, experiments are run in order to better understand how quickly the algorithm terminates. This is an important test to carry out before the experiments begin so as to make sure that enough chromosome generations are run in order for the solution to converge but, at the same time, an excessive number of generations will come at the expense of excessive computation time.

Figure 5.3 shows the convergence rate of the algorithm when the termination criterion is set to 400 iterations.

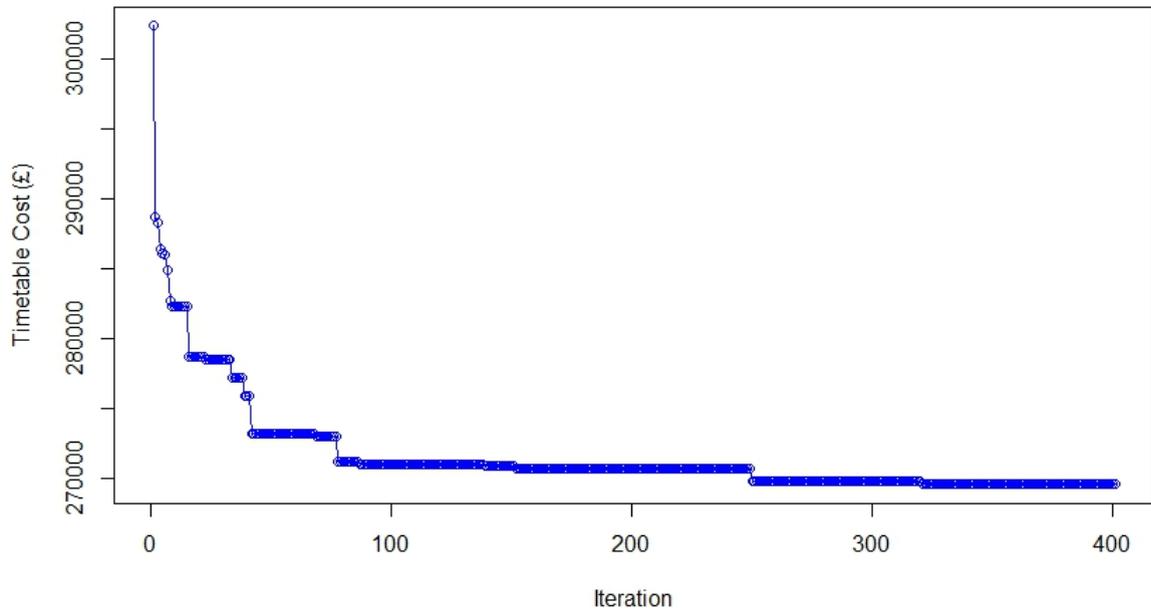


FIGURE 5.3: Convergence with 400 iterations

The figure above shows that, as expected, the largest reductions in the timetable cost occur during the initial iterations with later iterations leading to reductions of lower magnitude. Nonetheless, by the 100th iteration the algorithm seems to converge with only minimal reductions being observed after that point. The biggest decreases in the cost function seem take place before the 100th iteration with minimal improvements taking place after the 200th iteration. In more detail, no improvements more than 0.5% are observed after the 100th iteration. Consequently, it has been determined to terminate the algorithm after a maximum of 200 iterations which appears to offer a satisfactory balance between the quality of the solution and the time needed to produce it. However, if the network increases

in size or the time span examined extends, it is likely that more iterations will be needed before the algorithm converges.

5.4 Optimised sequence process

Experiments were carried out in order to understand whether the trains can be sequenced in a different way such that the timetable cost is reduced. This will be achieved by comparing the current sequence with which trains are scheduled in the existing timetable and then running the optimisation algorithm to examine whether a more efficient sequence can be found. This process will be undertaken for three different demand levels

- Low demand - Average train loading is set equal to 50% and corresponds to hours with little demand
- Average demand - Average train loading is set equal to 100% and corresponds to hours with moderate demand
- High demand - Average train loading is set equal to 130% and corresponds to hours with very high demand (e.g. morning and afternoon peak)

In order to isolate the impact from sequencing, no further trains will be added in the timetable.

5.4.1 Optimised sequence traits - Presentation of results

Figures 5.4, 5.5 and 5.6 present the results from the experiments and Table 5.3 summarises the results in the figures presented.

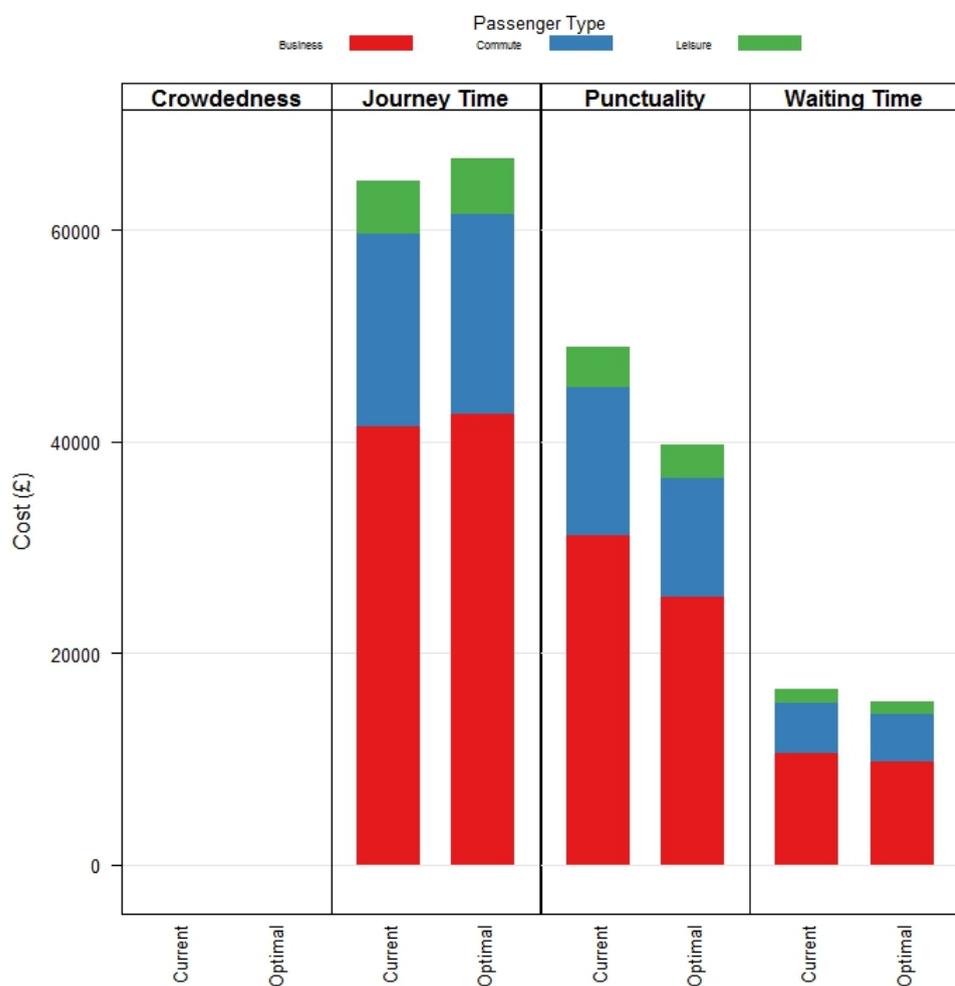


FIGURE 5.4: Optimised train sequencing for low demand levels

The results indicate that resequencing the trains can lead to lower timetable costs and this reduction is driven by reductions in the cost of punctuality. The cost of journey time and crowdedness increases slightly but this increase is mitigated by

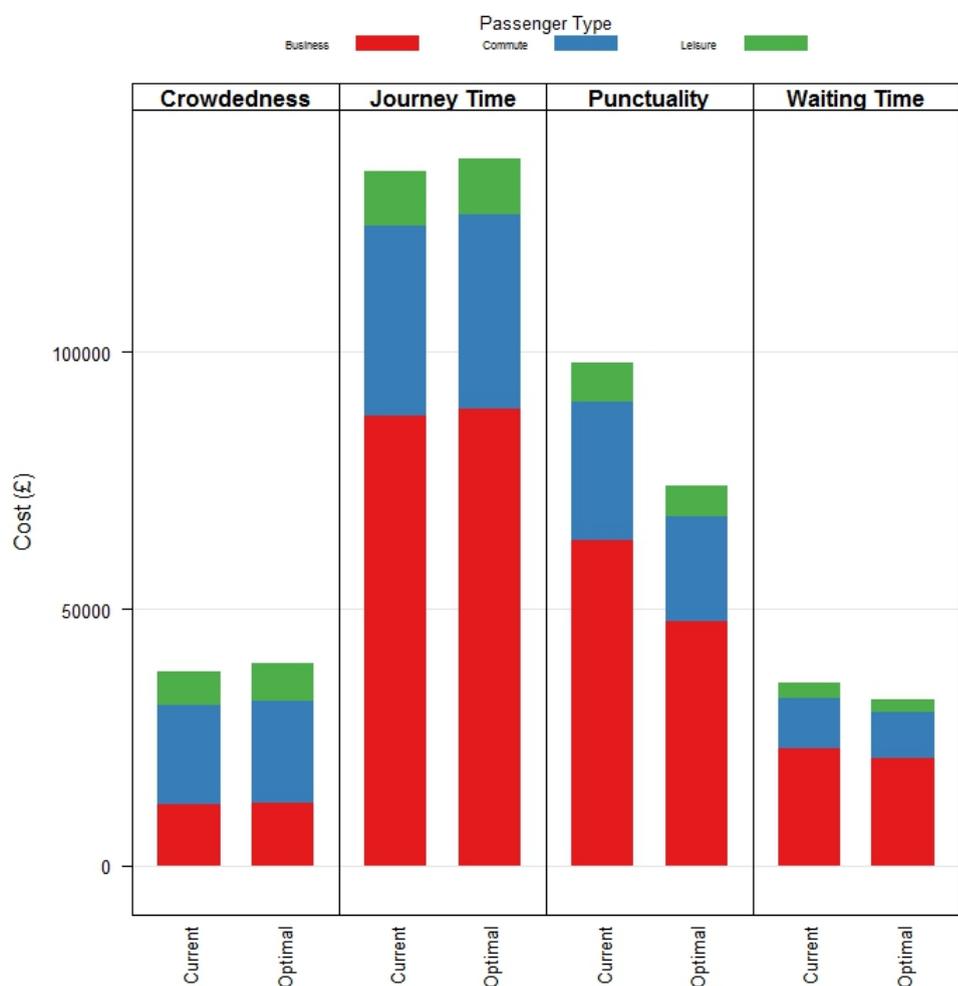


FIGURE 5.5: Optimised train sequencing for average demand levels

TABLE 5.3: Sequencing results for different demand levels

	Low Demand		Average Demand		High Demand	
	Current	Optimised	Current	Optimised	Current	Optimised
Crowdedness	0	0	37805	39386	138806	145998
Journey Time	64646	66689	135221	137477	176503	179386
Punctuality	48910	39664	97954	73815	140716	112679
Waiting Time	16576	15451	32458	32458	46207	42154

the significant improvements in the cost of punctuality¹. The results are consistent

¹The cost of crowdedness for the low demand scenario is zero due to the fact that no crowdedness penalty is charged for demand levels below 60%

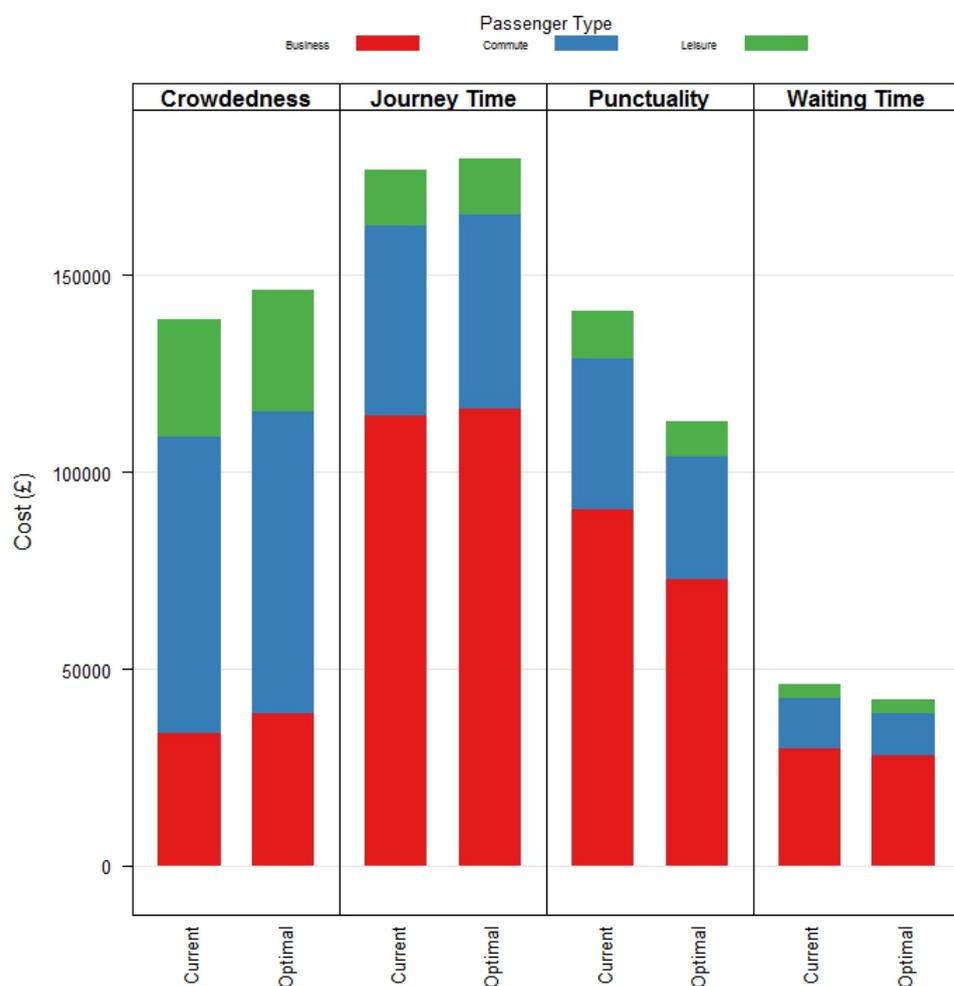


FIGURE 5.6: Optimised train sequencing for high demand levels

for all demand levels, implying that crowdedness levels do not affect the impact of sequencing.

5.4.2 Optimised sequencing traits - Discussion

The impact of reducing the timetable cost through sequencing can be understood by referring to the optimised sequence provided by the optimisation algorithm and

comparing it to the initial sequence. A closer look into the optimised sequence reveals that the reduction in punctuality comes from better distributing the trains which exit the network sooner.

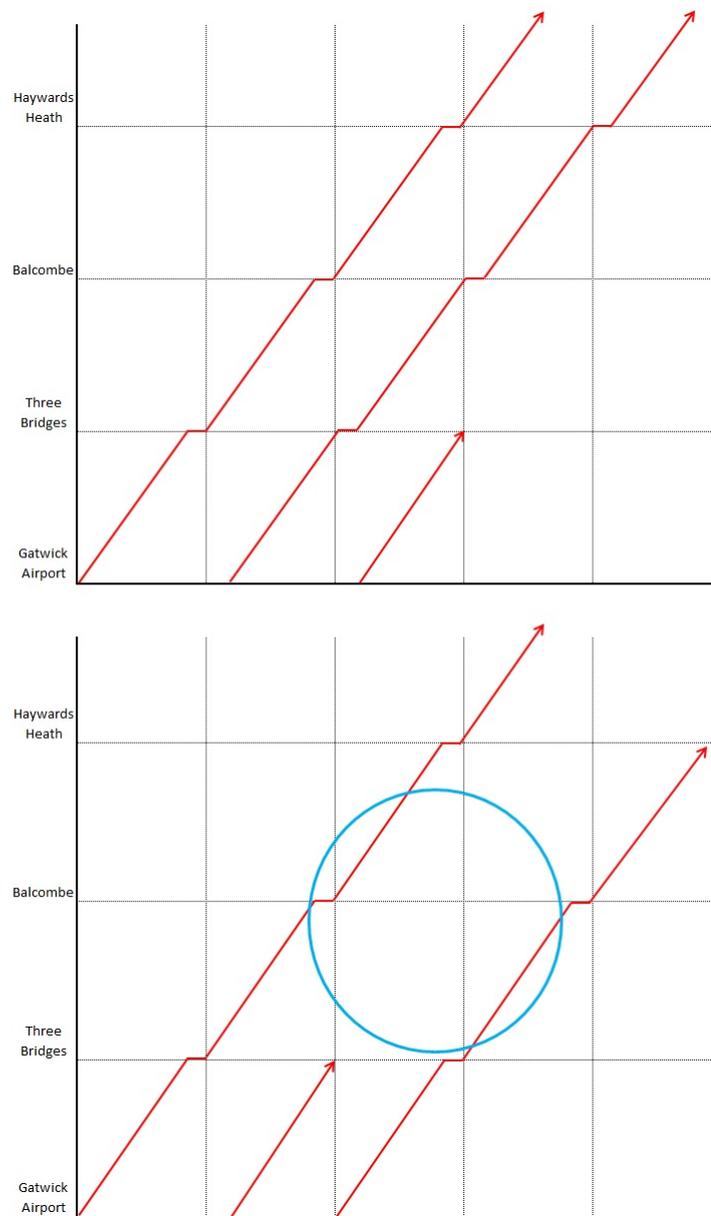


FIGURE 5.7: Optimised train sequencing illustration

Figure 5.7 illustrates an occasion in the current timetable where three trains are scheduled immediately after one another, leaving no time-buffer between trains

which can be used in case of a delay. This leads to higher punctuality costs since, if one train is delayed, the delay will inevitably be propagated onto other trains. The optimised sequence in 5.7 on the other hand shows that between the first and third train, a significant time gap exists as a result of the existence of a train which is scheduled in the time-gap between the above two trains but exits the network at Three Bridges instead of continuing all the way until Brighton. This serves the purpose of introducing time-buffers between the preceding and the succeeding train which can be used to absorb the buffer. Consequently, distributing the trains which exit the network early more evenly across the timetable contributes in reducing the impact of delays.

It should be emphasised that such a scheduling may not be possible in real life due to the fact that trains exiting Three Bridges will visit other stations in their path and this resequencing will affect the feasibility of the timetable as a whole. Boundary conditions nonetheless are bound to lead to certain inherent limitations in the problem due to the fact that only a subsection is considered rather than the network as a whole. However, the results shown in this section can provide a useful insight into the timetabling procedure when timetabling is carried out for the network as a whole.

5.5 Impact of scheduling additional trains

A series of experiments were carried out to gain an insight into how different crowdedness levels affect the optimal number of trains to be scheduled. As shown in Section 5.4, the optimised sequence traits remain the same irrespective of the number of passengers expected to utilise the services. However, crowdedness levels are expected to influence the optimised number of trains to be scheduled through the Hill Climbing heuristic.

For the purpose of this experiment, three demand levels will be examined which are the same as the ones used in Section 5.4

- Low demand - Average train loading is set equal to 50% and corresponds to hours with little demand
- Average demand - Average train loading is set equal to 100% and corresponds to hours with moderate demand
- High demand - Average train loading is set equal to 130% and corresponds to hours with very high demand (e.g. morning and afternoon peak).

In the low demand scenario, crowdedness levels are below the threshold for applying a penalty and, as such, no further trains are expected to be added since the increase in the cost of journey time and punctuality will offset any gains in terms of the waiting cost. In the average and high demand scenarios, trains are

expected to be added until the marginal improvements in crowdedness are offset by the increase in punctuality cost.

Each time the Hill Climbing heuristic adds a train, it does so by adding one train in each direction. Since in this case study there are two possible directions (Gatwick to Brighton and vice versa), each iteration of the heuristic adds a total of two trains. The extra trains to be added by the Hill Climbing heuristic are the express 375 classes which travel from Gatwick to Brighton (and vice versa) without stopping at any stations in between. The reason for adding express services rather than regional is the fact that the majority of passengers request a service for the specific destination (Appendix C), implying that faster trains can be added while serving a high proportion of the passengers at the same time. As mentioned in Section 5.2, in the current timetable there are 18 trains scheduled for the 'down' direction (Gatwick Airport to Brighton) and 22 trains scheduled in the 'up' direction (Brighton to Gatwick Airport).

5.5.1 Impact of overcrowding - Presentation of results

The results from the experiments are summarised in Figures 5.8, 5.9 and 5.10. The blue column indicates the optimised number of trains to schedule while the rightmost column in the figures (painted violet) indicates the point where the cost of scheduling an additional train overcomes the benefits.

The results show that, for low demand levels (Figure 5.8) adding one train in each

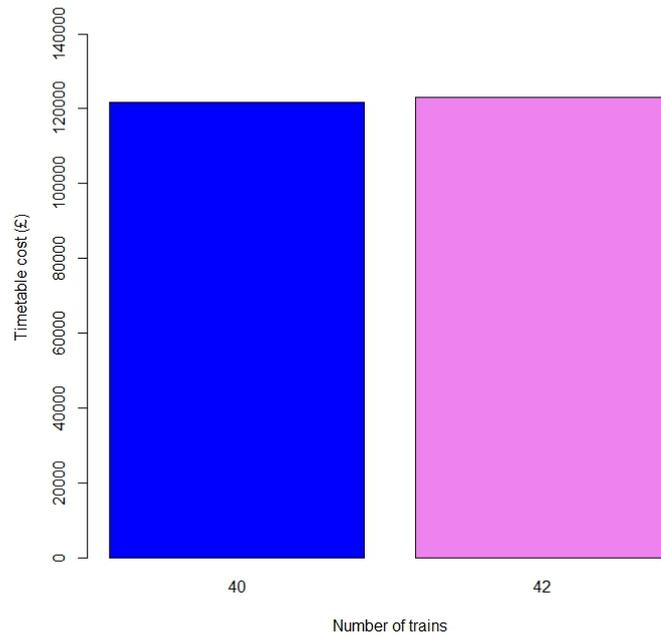


FIGURE 5.8: Optimised number of trains for low demand

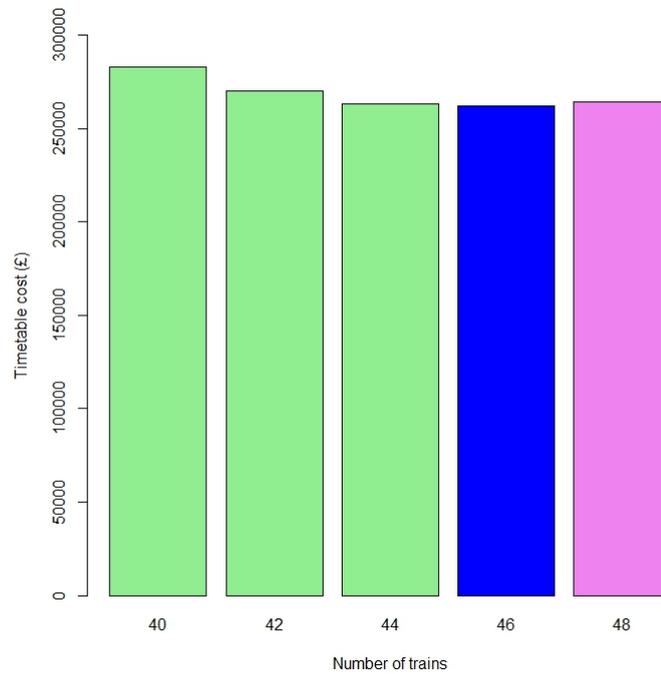


FIGURE 5.9: Optimised number of trains for average demand

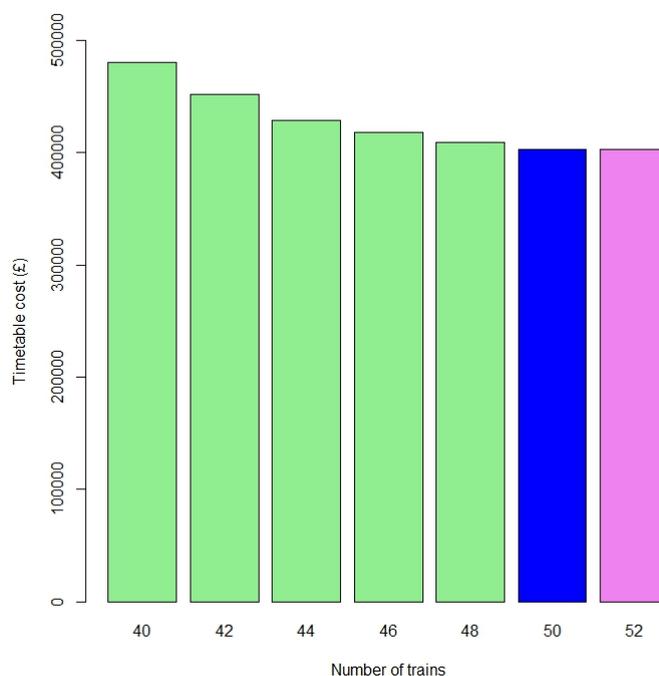


FIGURE 5.10: Optimised number of trains for high demand

direction causes the total timetable cost to rise and the optimisation procedure returns the optimised solution as the one which is comprised of 40 trains. The average demand scenario (Figure 5.9) allows for three additional trains to be added in each direction (six total) before the timetable cost rises. If the demand is raised even further (Figure 5.10), the total number of trains to be added is raised to five for each direction. It is important to note that if six further trains are added in each direction the timetable will be rendered infeasible due to the fact that the timetable's span will exceed two hours.

Figures 5.11, 5.12 and 5.13 show the marginal changes in timetable cost after the introduction of additional trains. It appears that, as expected, if the demand is low, scheduling additional trains leads to the timetable cost to be higher by

approximately 1200. When the demand is sufficiently high though, scheduling additional trains leads to lower timetable cost but the cost reductions are experiencing diminishing marginal return as the number of additional trains increases.

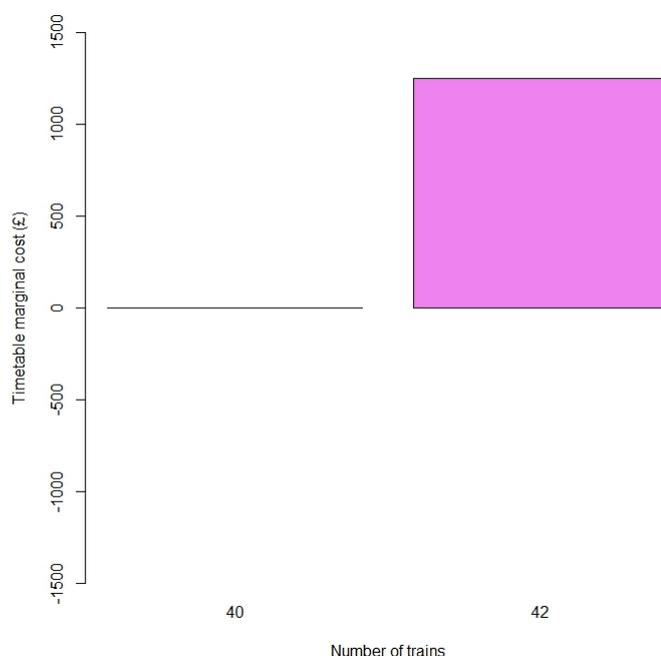


FIGURE 5.11: Marginal timetable improvements for the low demand scenario

In order to understand what cost functions are affected by the insertion of additional trains, a closer look is required on the value of each individual cost function. The results are summarised in Figures 5.14, 5.15, and 5.16 and Tables 5.4, 5.5 and 5.6 provides the cost of each function under each scenario.

TABLE 5.4: Optimised number of trains for low demand - Solution table

	Crowdedness	Journey Time	Punctuality	Waiting Time	Total cost
40 Trains	0	66689	39664	15451	121804
42Trains	0	66900	40975	15178	123053

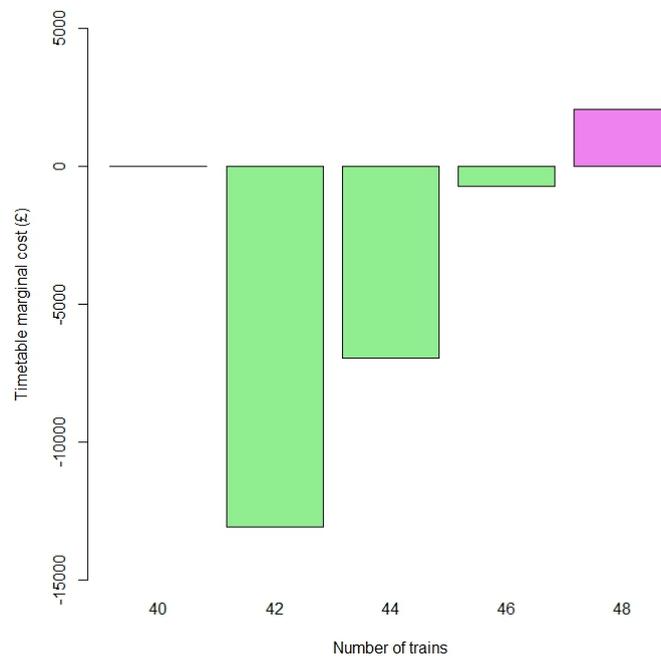


FIGURE 5.12: Marginal timetable improvements for the average demand scenario

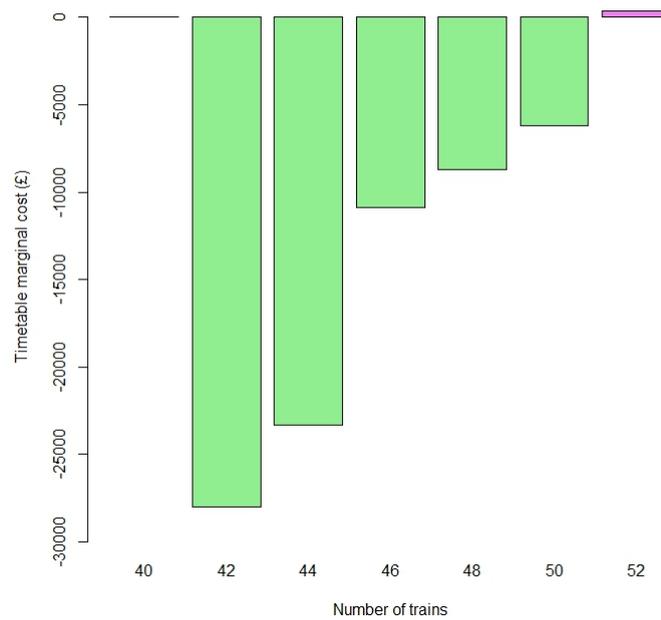


FIGURE 5.13: Marginal timetable improvements for the high demand scenario

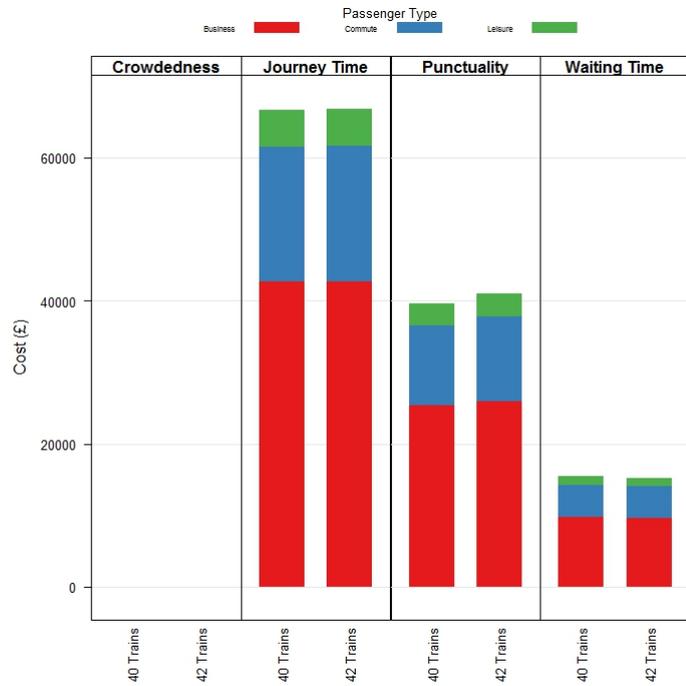


FIGURE 5.14: Optimised number of trains for low demand - Timetable break down

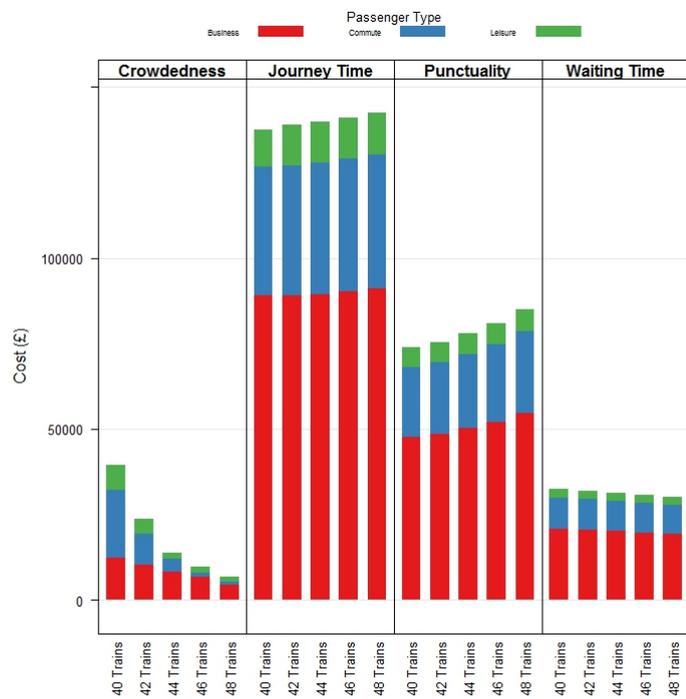


FIGURE 5.15: Optimised number of trains for average demand - Timetable break down

TABLE 5.5: Optimised number of trains for average demand - Solution table

	Crowdedness	Journey Time	Punctuality	Waiting Time	Total cost
40 Trains	39386	137477	73815	32458	283136
42 Trains	23758	138900	75446	31910	270014
44 Trains	13919	139786	78014	31371	263090
46 Trains	9564	141049	81021	30718	262352
48 Trains	6770	142448	85129	30089	264436

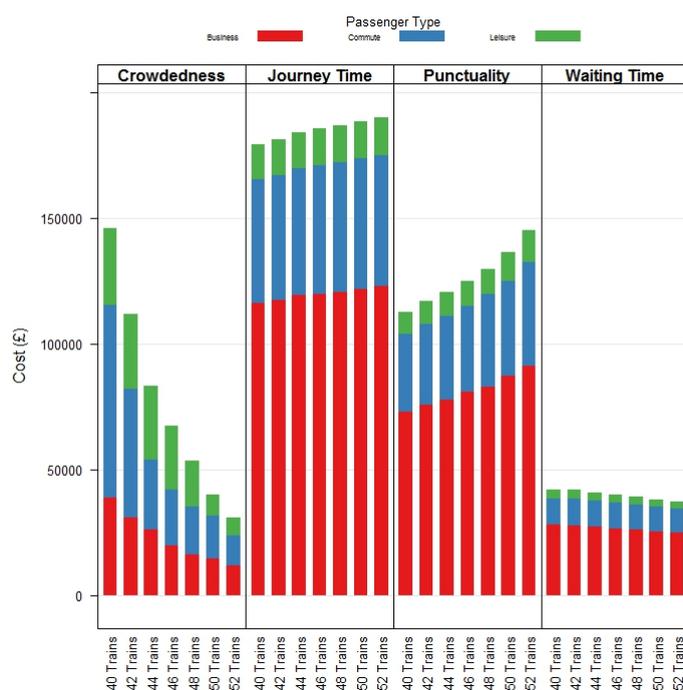


FIGURE 5.16: Optimised number of trains for high demand - Timetable break down

Breaking down the total timetable cost to the individual cost functions shows that the introduction of additional trains has an impact on all cost functions but the ones which are mostly affected are crowdedness and punctuality. This is also evident in Figures 5.17, 5.18, 5.19 and 5.20.

Introducing trains drives the cost of crowdedness down since more seats are offered

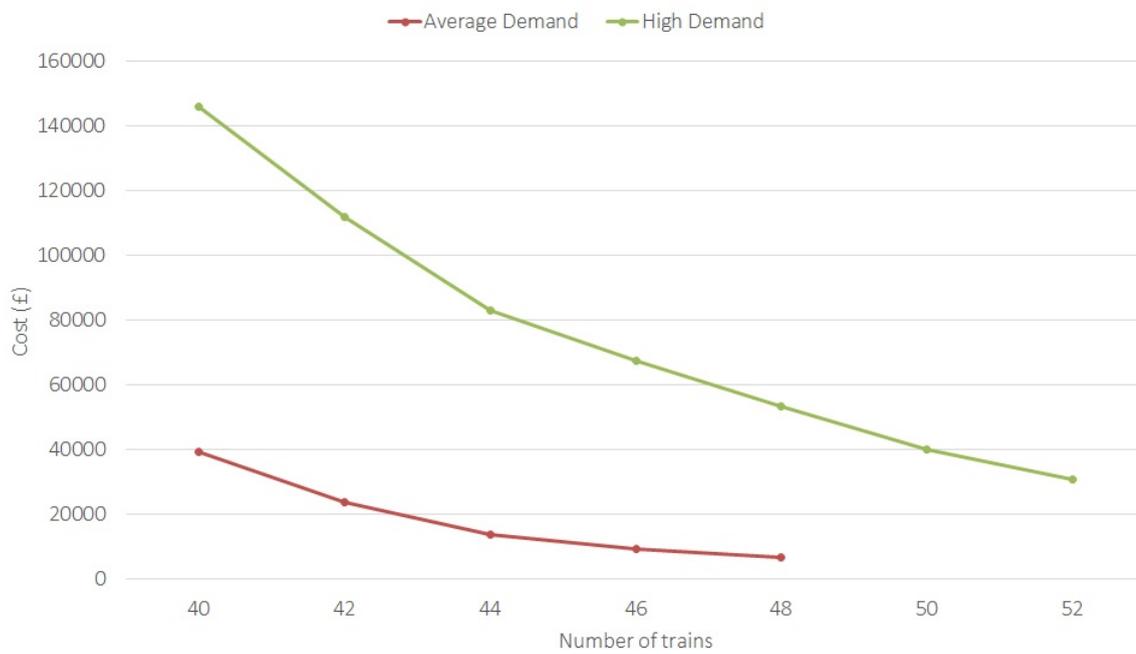


FIGURE 5.17: Changes in the cost of crowdedness when additional trains are scheduled

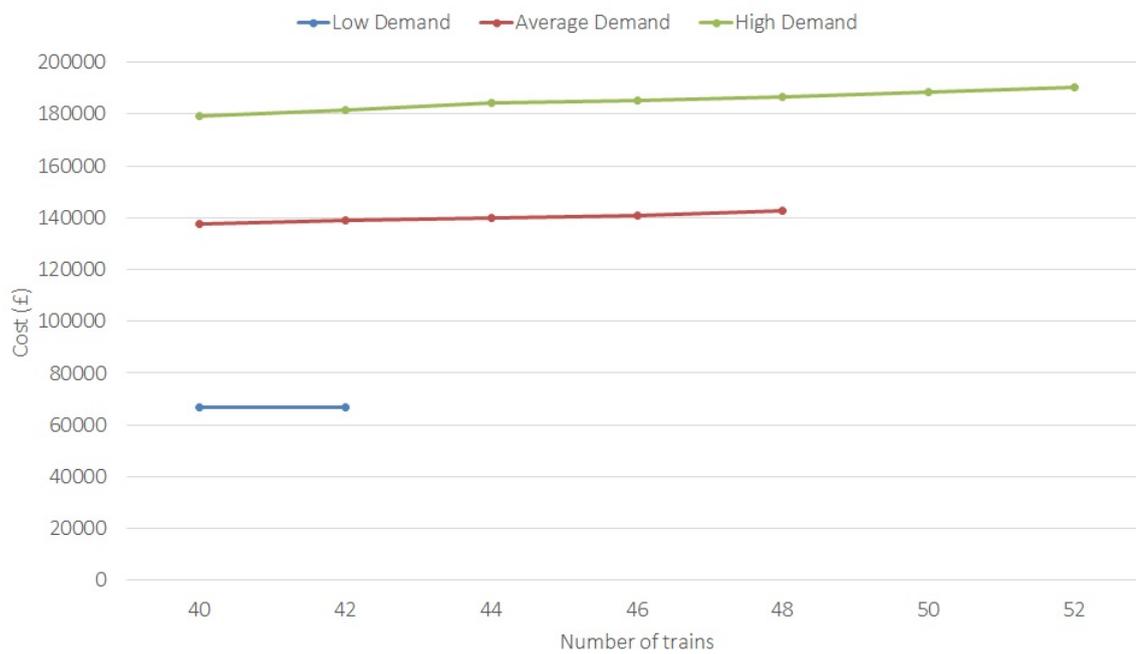


FIGURE 5.18: Changes in the cost of journey time when additional trains are scheduled

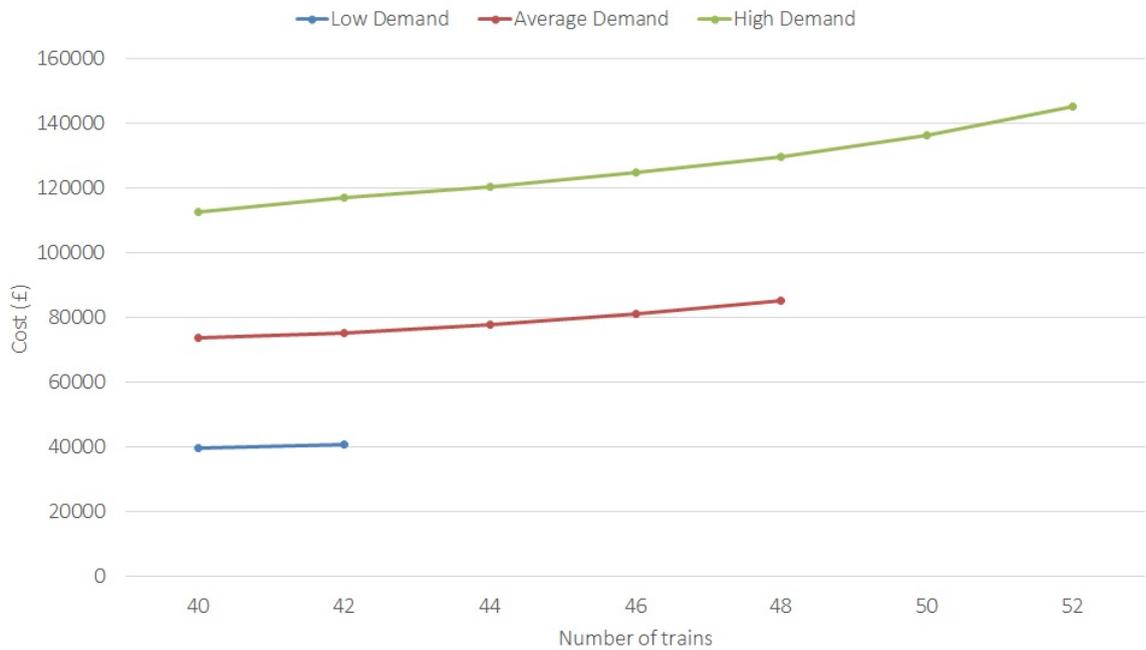


FIGURE 5.19: Changes in the cost of punctuality when additional trains are scheduled

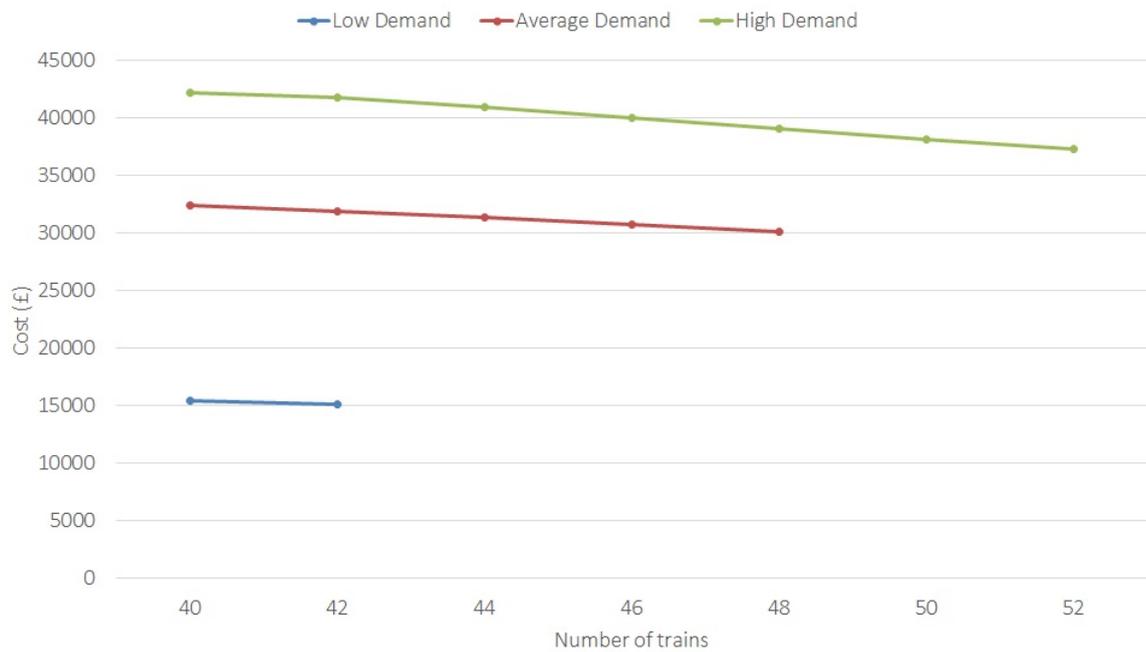


FIGURE 5.20: Changes in the cost of waiting time when additional trains are scheduled

TABLE 5.6: Optimised number of trains for high demand - Solution table

	Crowdedness	Journey Time	Punctuality	Waiting Time	Total cost
40 Trains	145998	179386	112679	42154	480217
42 Trains	112006	181404	116960	41800	452170
44 Trains	83182	184132	120581	40943	428838
46 Trains	67443	185423	125027	40035	417928
48 Trains	53609	186832	129753	39030	409224
50 Trains	40078	188457	136352	38146	403033
52 Trains	30931	190159	145053	37288	403431

between stations, leading to reduced crowdedness levels in the trains which is translated into a lower penalty. On the other hand, the scheduling of additional trains drives the cost of punctuality up due to the fact that more trains are now likely to be delayed and the delays have a knock-on effect on subsequent trains as well. With regards to the cost of journey time, the scheduling of additional trains causes more congestion in the bottleneck, leading to trains requiring more time to reach their destinations. This eventually translates into increases in the cost of journey time cost function as the number of trains increases. The cost of waiting time decreases slightly as the number of scheduled trains increases as there are more frequent trains to carry people to their destinations.

5.5.2 Impact of overcrowding - Discussion

A closer look at the results presented in Section 5.5 reveals that the scenario with low demand is a trivial solution to the issue of adding further trains while the solutions by the average and high demand scenarios present similar trains which

require further analysis to understand. These traits concern the behaviour of the cost of crowdedness and punctuality as well as the interaction between these two cost functions.

The scenario with low demand (Figure 5.14) shows that scheduling additional trains leads to increases in the cost of punctuality and the cost of journey time while the cost of waiting time decreases slightly. Crowdedness levels are below the threshold level of 60% meaning that the no penalty is applied for crowdedness. Consequently, the addition of one extra train has a negative impact on the timetable.

The solutions for the scenarios with average and high demand (Figure 5.15 and Figure 5.16) presents interesting results with respect to the impact of scheduling additional trains on the cost of crowdedness. The high demand scenario (Figure 5.16) shows that the scheduling the first two additional trains in each direction has a disproportionately high impact due to the fact that the addition of those trains helps to eliminate standing passengers who have extremely high penalties. It is interesting to note that, despite the fact that crowdedness penalties $R_{n,i \rightarrow j}$ in Equation 3.16 decrease linearly as crowdedness levels decrease (Table 3.3 and Table 3.4), the results in Figure 5.15 and Figure 5.16 show the timetable's crowdedness cost decaying non-linearly as the number of scheduled trains increases. This is attributed to the fact that the cost function given in Equation 3.16 is non-linear and also due to the way passengers are allocated to the trains by Equation 4.5.

When an additional train is scheduled, the equation's numerator remains constant but the denominator rises, resulting in trains experiencing a loading factor that decreases non-linearly. Consequently, each additional train leads to smaller reductions in the number of people on board each train, leading to the decreasing marginal benefits in the crowdedness cost shown in Figures 5.15 and 5.16. However, the total journey time increases which means that the trains may be less crowded but passengers spend more time inside the trains which leads to the reductions in the cost of crowdedness to be slightly mitigated by the increase in running times. The increase in cost of journey time is caused by the fact that trains need more time to travel from their origin to destination because of congestion in the bottleneck in Keymer Junction.

Punctuality is influenced by the scheduling of additional trains in a more straightforward way. With the addition of each extra train, the number of trains likely to be delayed increases and since the delay of a train has a knock-on effect on subsequent trains, the timetable's cost of punctuality increases exponentially as the number of scheduled trains increases (Figures 5.15 and 5.16). These results are consistent with the findings of previous authors such as Gibson et al. [38] who showed that impact of punctuality increases exponentially as the railway network gets more congested.

The behaviour of the crowdedness and punctuality cost functions leads to an important trade-off which governs the optimal number of trains to be scheduled given the different demand levels. The marginal cost is always bound to exceed

the marginal benefits of crowdedness through the scheduling of more trains. This occurs due to the fact that as extra trains are scheduled, the marginal losses from punctuality increase exponentially while the marginal benefits of crowdedness decrease. As illustrated in Figures 5.14, 5.15 and 5.16, the equilibrium point depends on the levels of crowdedness. The number of trains which constitute part of the optimised solution changes depending on whether there is low, average or high demand.

5.6 Pareto analysis

Analysing the different trade-offs also involves examining the Pareto Frontier to better understand how the optimised solution to the problem changes when the cost function coefficients vary. The Pareto Frontier for two objective functions can be represented by a curve where each point on the curve indicates an efficient solution when the two objective functions are being optimised simultaneously [33].

The objective function in Equation 4.1 is constructed by adding all the individual cost functions which comprise it. The Pareto Frontier will be constructed by multiplying each cost function by a scalar a_q such that:

$$\min_{\tau, \sigma} : C = a_1 C_T + a_2 C_W + a_3 C_P + a_4 C_D \quad \sum_{q=1}^4 a_q = 1, \quad (5.1)$$

Equation 5.1 will enable for the construction of the frontier which will represent

all the solutions that are Pareto efficient since, as Ehrgott [33] states, all optimised solutions of scalarised problems are always Pareto efficient.

According to Equation 5.2, eleven different combinations could be selected to use in the Pareto analysis.

$$\frac{4!}{2!2!} + \frac{4!}{3!1!} + \frac{4!}{4!0!} \quad (5.2)$$

However, it was decided that analysing all combinations was not possible due to the time required to run the experiments needed to construct the Pareto Frontier.

Consequently, only three combinations were used in the Pareto analysis:

- Crowdedness against punctuality
- Crowdedness against journey time
- Punctuality against journey time

These three combinations were chosen since, together they have the largest contribution towards the timetable's total cost. This is evident by referring to Tables 5.3, 5.4, 5.5 and 5.6 where the cost of waiting has the smallest impact out of all four cost functions. Even though it is understood that all possible trade-offs need to be analysed, the time constraints imposed by the project necessitates that the focus be shifted on the combinations which are the most likely to provide the most useful insight.

The Pareto Frontier was determined by solving the optimisation problem using different values for the scalars a_q . Sections 5.4 and 5.5 have shown that changes in the value of the Crowdedness and Punctuality cost functions dominate the decisions made to derive the optimised timetable. Following this, experiments were aimed at constructing the Pareto Frontier for the Crowdedness and Punctuality cost function. Table 5.7 shows the different values of a_q used to construct the Pareto Frontier for the Crowdedness and Punctuality cost functions. The Journey Time and Waiting Time cost functions which are not included in the experiment have the parameters set equal to zero (i.e. $a_1 = 0$ and $a_2 = 0$) so as not to interfere with the Pareto analysis of the Crowdedness and Punctuality cost functions.

TABLE 5.7: Pareto Frontier construction for Crowdedness and Punctuality

	Crowdedness	Punctuality
Scenario 1	0.00	1.00
Scenario 2	0.15	0.85
Scenario 3	0.25	0.75
Scenario 4	0.35	0.65
Scenario 5	0.40	0.60
Scenario 6	0.50	0.50
Scenario 7	0.60	0.40
Scenario 8	0.65	0.35
Scenario 9	0.70	0.30
Scenario 10	0.75	0.25
Scenario 11	0.85	0.15
Scenario 12	1.00	0.00

Experiments only focused on the scenarios with average and high demand since, in the low demand scenarios the cost of crowdedness is zero making the low demand scenario trivial.

The trade-off between the Crowdedness and Journey Time cost function determines the optimised solution in cases where the marginal changes in punctuality are lower than the marginal changes in the journey time. Table 5.8 shows the different values of a_q used to construct the Pareto Frontier for the Crowdedness and Journey Time cost functions. The parameters for the Punctuality and Waiting Time cost functions (i.e. a_3 and a_2 respectively) are set equal to zero.

TABLE 5.8: Pareto Frontier construction for Crowdedness and Journey Time

	Crowdedness	Journey Time
Scenario 1	0.00	1.00
Scenario 2	0.15	0.85
Scenario 3	0.25	0.75
Scenario 4	0.35	0.65
Scenario 5	0.40	0.60
Scenario 6	0.50	0.50
Scenario 7	0.60	0.40
Scenario 8	0.65	0.35
Scenario 9	0.70	0.30
Scenario 10	0.75	0.25
Scenario 11	0.85	0.15
Scenario 12	1.00	0.00

As mentioned before, the low demand scenario was excluded from the analysis since the interaction of the Crowdedness and Journey Time cost functions is of interest only the cost of the Crowdedness cost function is not zero.

The Pareto Frontier for the Punctuality and Journey Time cost functions was constructed since the Journey Time cost function has a cost of high magnitude (see Tables 5.4, 5.5 and 5.6), with the potential to have a significant impact during the optimisation process. This is more evident in the low demand scenario where

the interaction between Punctuality and Journey time determines the optimised solution. Table 5.9 shows the different values of a_q used to construct the Pareto Frontier for the Punctuality and Journey Time cost functions. Once again the parameters for the cost functions not included in the experiments (namely Crowdedness and Waiting Time with parameters a_4 and a_2 respectively) are set equal to zero.

TABLE 5.9: Pareto Frontier construction for Punctuality and Journey Time

	Punctuality	Journey Time
Scenario 1	0.00	1.00
Scenario 2	0.10	0.90
Scenario 3	0.15	0.85
Scenario 4	0.25	0.75
Scenario 5	0.35	0.65
Scenario 6	0.40	0.60
Scenario 7	0.50	0.50
Scenario 8	0.60	0.40
Scenario 9	0.65	0.35
Scenario 10	0.70	0.30
Scenario 11	0.85	0.15
Scenario 12	1.00	0.00

In total, seven different Pareto Frontiers will be summarised in Table 5.10.

TABLE 5.10: Pareto Frontier to be constructed

Cost function 1	Cost function 2	Demand levels		
		Low	Average	High
Crowdedness	Punctuality		*	*
Crowdedness	Journey Time		*	*
Punctuality	Journey Time	*	*	*

When the experiments for the seven scenarios in Table 5.10 are run, results will be plotted on a scatter plot and a line will be fitted to understand the mathematical relationship governing the relationship of the variables in the plot.

The Waiting Time cost function was omitted from the Pareto analysis due to the fact that waiting time has the lowest magnitude of all the cost functions and it also experiences the smallest marginal changes when the trains are resequenced and when additional trains are being scheduled.

5.6.1 Pareto analysis - Presentation of results

The results shown in this section, illustrate the Pareto Frontiers which examine the relationships identified in Table 5.10. On top of each figure, the equation of the curve fitting the points is given.

The Pareto Frontier for Crowdedness against Punctuality can be seen in Figure 5.21 for average demand levels and Figure 5.22 for high demand levels.

Points located at the bottom-right of the plots indicate the scenarios where punctuality had a much bigger weight than crowdedness. As the weight of crowdedness increases, the cost of crowdedness also decreases but this improvement comes at the expense of an exponentially increasing cost of punctuality. Both graphs show that a logarithmic relationship governs the dynamic between the Crowdedness

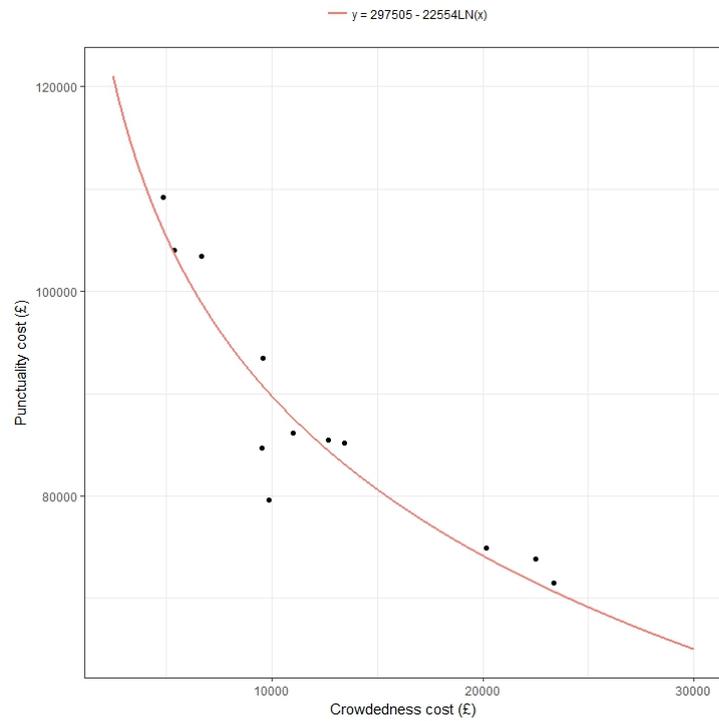


FIGURE 5.21: Pareto frontier for crowdedness against punctuality - Average demand

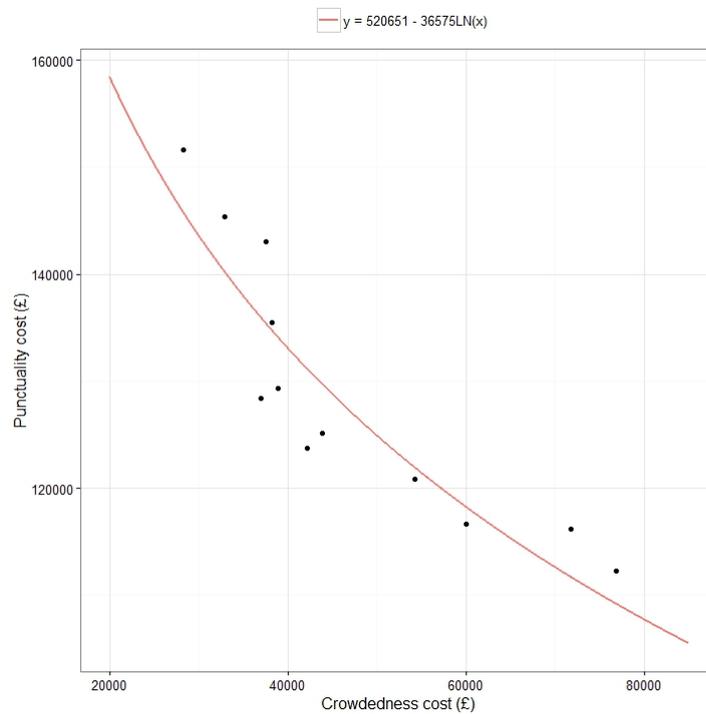


FIGURE 5.22: Pareto frontier for crowdedness against punctuality - High demand

and Punctuality cost functions. This implies that, as the cost of crowdedness increases, the cost of punctuality decreases but the marginal improvements decrease at a logarithmic rate. For example, consider Figure 5.21 which illustrates that, under average demand, the Pareto Frontier is expressed as

$$y = 297505 - 22554 \ln x \quad (5.3)$$

Equation 5.3 indicates that when Punctuality has a coefficient $a_p = 1$ and Crowdedness a coefficient $a_c = 0$, the optimisation algorithm will only focus on minimising the cost of punctuality by setting Equation 5.3 equal to zero. This is achieved when Crowdedness has a cost of approximately £534988². Similarly, when Punctuality has a coefficient $a_p = 0$ and Crowdedness a coefficient $a_c = 1$, crowdedness will be minimised when the cost of punctuality is set to an arbitrarily high number³

Figures 5.23 and 5.24 depict the interaction between the Crowdedness and Journey Time cost functions. The results from these experiments fall closer to the Pareto Frontier compared to the results in Figures 5.21 and 5.22 due to the fact that both Crowdedness and Journey Time are deterministic, decreasing the variability in the plot.

Similar to Figures 5.21 and 5.22, the relationship between these two objective functions can be represented using a logarithmic curve. In addition, akin to the

²In practice, the cost of punctuality cannot be completely eradicated due to the presence of primary delays

³This value is not necessarily feasible since, if enough trains are added, the timetable span may exceed the threshold, setting a lower bound to the minimum cost of crowdedness

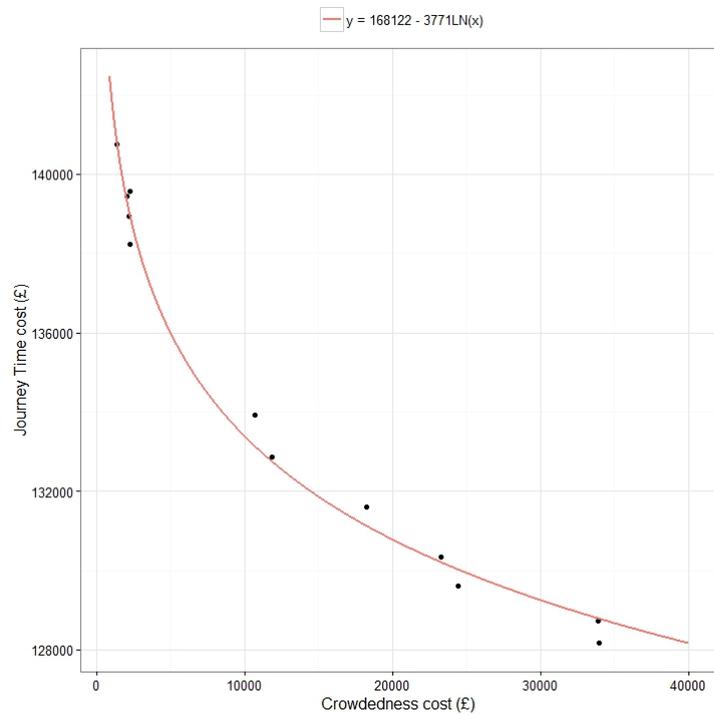


FIGURE 5.23: Pareto frontier for crowdedness against journey time - Average demand

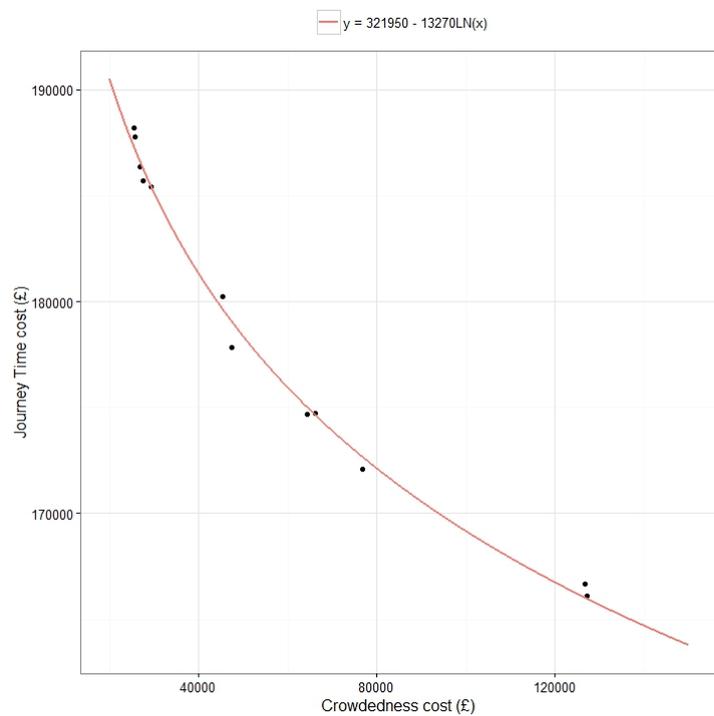


FIGURE 5.24: Pareto frontier for crowdedness against journey time - High demand

Crowdedness against Punctuality frontier, the extremes of the curves fitted in Figures 5.23 and 5.24 may not necessarily be feasible. This is attributed to the fact that, as long as the timetable consists of a single train, the cost of Journey Time will never be zero while Crowdedness may be prevented from being set to zero due to the constraints concerning the span of the timetable⁴.

The final Pareto Frontiers will analyse the interaction between Punctuality and Journey Time and the results are summarised in Figures 5.25, 5.26 and 5.27.

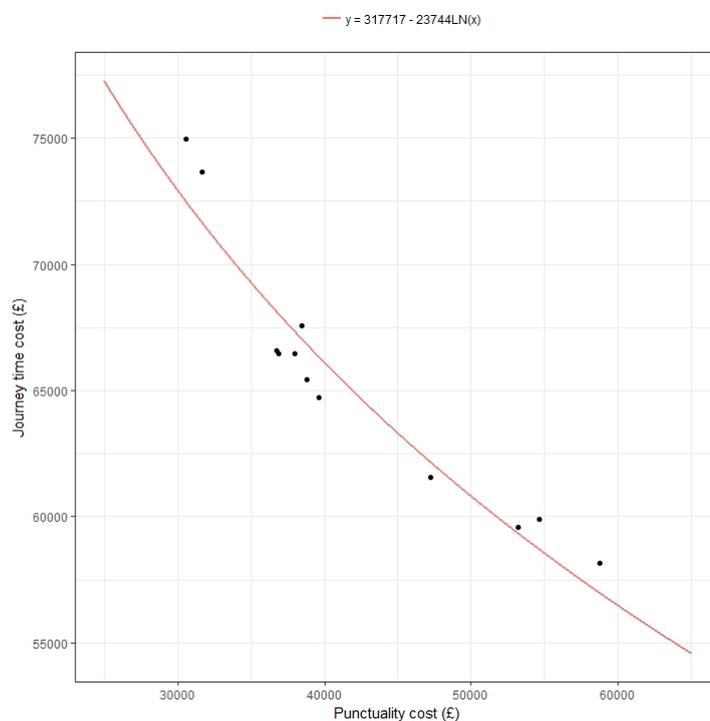


FIGURE 5.25: Pareto frontier for punctuality against journey time - Low demand

The low demand and average demand case, indicate that a logarithmic curve best fits the data. The high demand case is slightly different since a second order

⁴See footnote 3

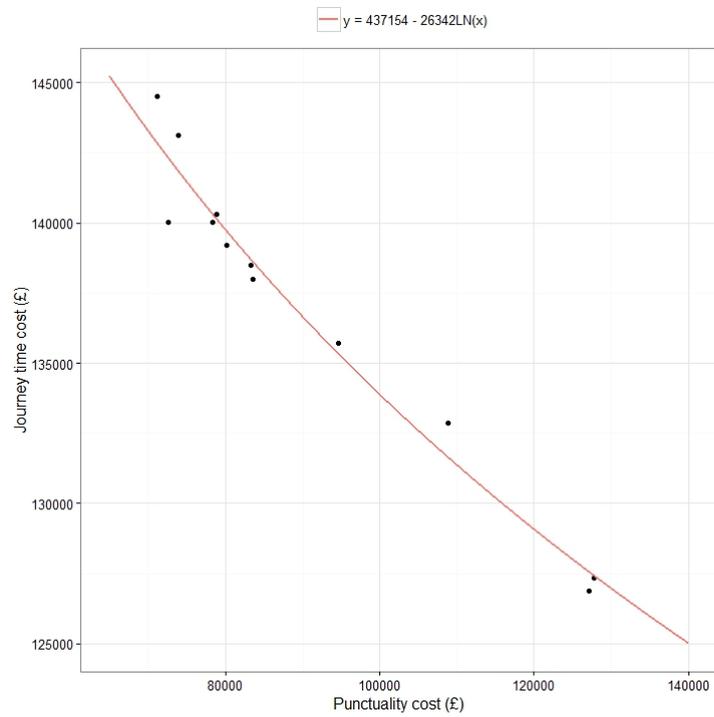


FIGURE 5.26: Pareto frontier for punctuality against journey time - Average demand

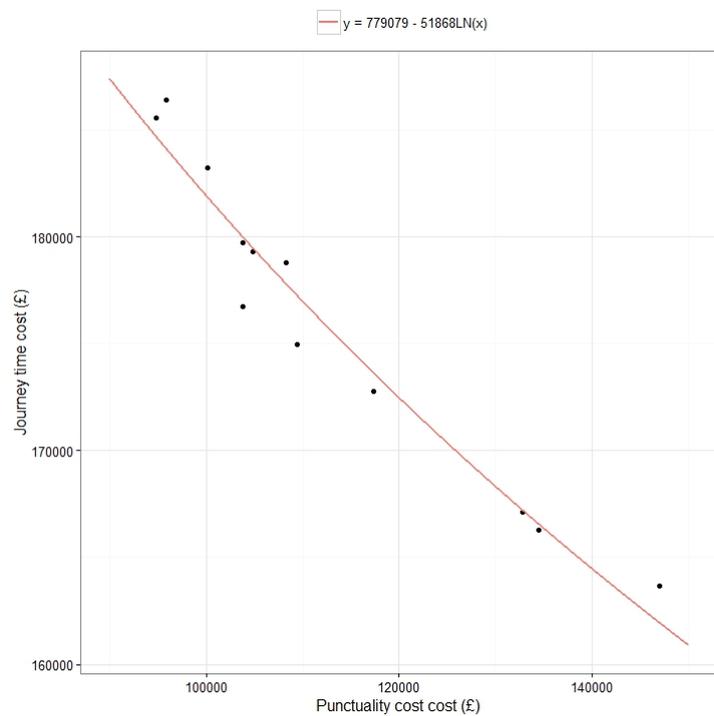


FIGURE 5.27: Pareto frontier for punctuality against journey time - High demand

polynomial curve best fits the data with an R^2 value

$$R_{Low}^2 = 0.972 \quad (5.4)$$

$$R_{Average}^2 = 0.962 \quad (5.5)$$

$$R_{High}^2 = 0.971 \quad (5.6)$$

Nonetheless, a logarithmic curve is, from a conceptual point of view, a more appropriate fit, leading to the attempt to fit a logarithmic curve to the data. The results obtained indicate that the logarithmic curve is an equally good fit with an R^2 value

$$R_{Low}^2 = 0.921 \quad (5.7)$$

$$R_{Average}^2 = 0.964 \quad (5.8)$$

$$R_{High}^2 = 0.953 \quad (5.9)$$

Due to the minimal differences in terms of the R^2 of the curves fitting the data, it was decided to keep the logarithmic curve.

Consequently, for all the Pareto Frontiers constructed (Table 5.10), a logarithmic curve offers the best mathematical representation for the Pareto Frontier. As demonstrated though, the extreme points of the curve may not always be attainable in practise.

5.6.2 Pareto analysis - Discussion

A detailed discussion of the results in Section 5.6.1 is provided in this section so as to get a more thorough of the Pareto Frontiers governing the interaction of the Crowdedness, Punctuality and Journey Time cost functions.

The first thing that becomes apparent from the results is that, certain plots exhibit less variability than others. For example, Figures 5.21, 5.22, 5.25, 5.26 and 5.27 have data points lying further away from the frontier compared to Figures 5.23 and 5.24. In theory, all of the optimisation results should lie on the Pareto Frontier but, in practice, this is unlikely to happen due to two reasons. The first one is that the optimisation algorithm may not necessarily return the optimal solution but a solution which is good enough; leading to the result being present close to the frontier but not on it directly. The second reason concerns the fact that one of the objectives plotted is Punctuality which is determined in a stochastic way, as explained in Section 4.4. This implies that, in every realisation of the timetable, different trains will be delayed and by a different amount which will leads to inherent inconsistencies that become apparent when the results are plotted.

The second thing to note, is the difference in the slope of the figures and its relation to the different demand levels. For example, Figure 5.21 describes the Pareto Frontier by plotting the cost of Crowdedness on the x-axis ($C_{D_{av}}$) against the cost of Punctuality on the y-axis ($C_{P_{av}}$) for average demand levels. The slope

of Pareto Frontier is

$$C_{P_{av}} = 297505 - 22554 \ln C_{D_{av}} \quad (5.10)$$

$$\frac{d}{dC_{D_{av}}}(297505 - 22554 \ln C_{D_{av}}) \quad (5.11)$$

$$\frac{dC_{P_{av}}}{dC_{D_{av}}} = -\frac{22554}{C_{D_{av}}} \quad (5.12)$$

for Crowdedness levels approximately in the range

$$5000 < C_{D_{av}} < 25000 \quad (5.13)$$

Calculating the slope of the frontier for the range provided gives

$$-4.51 < \frac{dC_{P_{av}}}{dC_{D_{av}}} < -0.9 \quad 5000 < C_{D_{av}} < 25000 \quad (5.14)$$

Extending this to Figure 5.22 which constructs the Pareto Frontier for Crowdedness ($C_{D_{hi}}$) against Punctuality ($C_{P_{hi}}$) for high demand results in

$$C_{P_{hi}} = 520651 - 36575 \ln C_{D_{hi}} \quad (5.15)$$

$$\frac{d}{dC_{D_{hi}}}(520651 - 36575 \ln C_{D_{hi}}) \quad (5.16)$$

$$\frac{dC_{P_{hi}}}{dC_{D_{hi}}} = -\frac{36575}{C_{D_{hi}}} \quad (5.17)$$

which leads to the slope

$$-2.44 < \frac{dC_{P_{hi}}}{dC_{D_{hi}}} < -0.46 \quad 15000 < C_{D_{hi}} < 80000 \quad (5.18)$$

It is therefore apparent that the frontier for average demand levels has, on average, a steeper slope (Equation 5.14) compared to the frontier for high demand levels (Equation 5.18). This makes intuitive sense since, when crowdedness is being multiplied by a scalar a_q close to zero, no extra trains will be added, leading to the timetable having a huge Crowdedness cost. However, as the scalar a_q for crowdedness increases, more trains are likely to be added. In the average demand scenario, this means that initial improvements in the cost of Crowdedness will be achieved without sacrificing the timetable's cost of Punctuality. As the Crowdedness scalar moves closer to one, the curve will quickly steepen due to the fact that the improvements in the cost of Crowdedness come at the expense of significant losses in the cost of punctuality. In the high demand scenario though, the trains are so overcrowded that, even when the scalar for Crowdedness is close to one, significant improvements in terms of the cost of Crowdedness can be achieved without the cost reductions being offset by the cost of Punctuality. These findings can also be extended to Figures 5.23 and 5.24 which also have Crowdedness on the x-axis.

5.7 Summary

This chapter has used the formulations of the cost functions provided in Chapter 3 and the optimisation algorithm in Chapter 4 to carry out experiments which provided an insight into different timetabling parameters and the summary of

those experiments is provided in this section.

Section 5.2 outlines the Gatwick Airport to Brighton section of the Brighton Main Line which is the network used for the experiments. The timetable considered covers the morning peak time period 08:00-10:00 during which two different train types traverse the network and the passenger mix is comprised of 20% business and leisure passengers and 60% commuting passengers.

Following the description of the network, experiments were carried out in Section 5.3 to determine how quickly the optimisation algorithm converges for the given problem. It has therefore been determined to run 200 iterations since the specific number of runs was striking an acceptable balance between the solution quality and computation time.

Section 5.4 signals the beginning of the main bulk of the experiments by examining whether trains can be sequenced in such a way that the cost of the timetable can be reduced through resequencing only. The results have shown that, when the trains scheduled to exit the network quickly are distributed more evenly across the timetable, then an artificial buffer is inserted which can absorb delays, reducing the cost of a timetable. This is a trait which remains irrespective of whether the loading factor of the trains.

While Section 5.4 focuses purely on train sequencing, Section 5.5 examines the effect of scheduling additional trains. In general, the number of additional trains to be scheduled depends on the levels of loading factor of the trains, with timetables

constructed under heavy demand scenarios being the ones that benefit the most from the introduction of additional trains. Furthermore, the optimised number of additional trains is mainly depended on the equilibrium between the Crowdedness and Punctuality cost functions. The reason for this is because as the number of scheduled trains increases, the timetable cost of crowdedness is showing traits of marginal diminishing returns while the cost of punctuality increases exponentially.

Finally, Section 5.6 explains how a series of experiments is run in order to construct the Pareto Frontier for a number of cost function combinations. A logarithmic curve has been decided to represent to represent the frontier since, not only provides the best fit for the data, but is also meaningful from a conceptual point of view. When Crowdedness is considered in the analysis, it has been shown that the Pareto Frontier becomes flatter for the range of crowdedness values the experiments were carried out. This has been shown to be due to the high crowdedness penalties associated with high demand levels and the ability of the timetable to incorporate more trains before the reductions in terms of the cost of crowdedness become trivial.

Chapter 6

Conclusions

6.1 Thesis overview

Chapter 1 of the thesis introduces the British railway industry and the describes timetabling process that is currently being undertaken. As of the time of writing, there is no way of calculating a objective value for the performance of a railway timetable. Consequently, there is no way to objectively compare two timetables in order to determine which one is better. This creates the need to formulate a set of objective functions which can systematically evaluate a railway timetable to determine how good it is. This set of the objectives will be based on the framework created by Chen and Roberts [20]. In case of formulating more than one objective, an analysis should be carried out to examine how these objectives interact under different problem parameters.

The first part of Chapter 2, reviews the literature on railway timetabling. Even though extensive literature exists which formulates the train timetabling problem as a multi-objective optimisation problem, most of the papers only use two formulations. However, a small number of authors optimise more than two objectives but these objectives measure punctuality using different formulations, resulting to a limited breadth of analysis. Furthermore, little to no attempt is being made to analyse the interaction between the objective functions the authors are using. This arises from the fact that the vast majority of the literature focuses on the algorithm used to solve the problem instead of provided an extensive analysis of the cost functions used. An exception to this is the case where punctuality and network capacity are optimised in which case a significant amount of literature has been devoted to explaining their relationship. The latter part of the first section in Chapter 2 is devoted to the different formulations developed to examine performance metrics related to capacity (both network and system capacity), journey time, punctuality and waiting time.

Chapter 2 concludes by describing the different optimisation algorithms developed over the years to tackle the train scheduling problem. The fact that the problem is computationally intractable leads to the use of heuristics and meta-heuristics to obtain approximate solutions. Some of the algorithms being used are Branch and Bound algorithms, Genetic algorithms and sub-gradient optimisation algorithms.

In Chapter 3 the specification of a railway timetable is defined which defines the variables used in the formulation of the cost functions as well as the constraints

needed to formulate feasible timetables. The cost functions formulated evaluate a timetable's non-monetary cost by examining its performance in terms of crowdedness, journey time, punctuality and waiting time. Crowdedness calculates the time penalty for passengers who travel in congested trains and Journey Time calculates the time it takes to travel from the origin to the destination of all passengers. Punctuality is measured as the time deviation of a train's expected arrival time from its scheduled arrival time at a station. Waiting Time penalises the time that customers have to wait before their service arrives. Monetary costs (e.g. the price that operators pay to buy the franchise, the operating costs of running a train etc.) were not considered as they are considered strictly confidential information and is not disclosed to the public.

It is immediately obvious that the objective functions have different dimensions and, in order to be combined, they must be adjusted such that they have the same dimension. The concept of travel time savings (also known as value of time) is therefore introduced which assigns monetary costs to different actions related to travelling. Travel time valuations are split into monetary and non-monetary costs with monetary cost covering costs which are being paid by the passenger (e.g. the cost of purchasing a ticket etc.) and non-monetary cost consider the opportunity cost of travelling in monetary terms (e.g. the opportunity cost of travelling, arriving late etc.). Travel time valuations has been used over the years by the Department for Transport to evaluate the benefits of investments in transport. This means that, up to now, travel time valuations are used more on a strategic rather than an operational level. Applying the cost of travel time savings to each

of the objectives we formulate, achieves the purpose of making each cost function measure the monetary cost of each objective function. Consequently, adding the objective functions together calculates a timetable's total non-monetary cost.

An optimisation algorithm is then presented in Chapter 4 which will enable for the analysis of the cost functions to take place. The algorithm works in three stages with a Genetic Algorithm in the first stage which evaluates the different sequence with which trains can be dispatched from their origin. Following the construction of a sequence, Dijkstras Algorithm is run to determine the shortest path from the origin to the destination of each train to be scheduled. In case of a clash between two trains at an node, priority is given to the train which appears first in the sequence given by the Genetic Algorithm. The final stage in the algorithm run a Hill-Climbing Heuristic which adds trains in the timetable and stops doing so when the extra train either increases the cost of the timetable or causes the timetable's time-span to exceed a given threshold.

Collecting information about the arrival rate of customers requesting a service for each origin-destination is made impossible due to the confidentiality agreements protecting such data. An alternative methodology is therefore described in the second part of Chapter 4 which works by constructing an origin-destination matrix each entry of which represents the origin and each row the destination station. The entries of the matrix contain the proportion of passengers (as a function of the train's total seats) who wish to utilise the specific origin-destination.

The BRaVE simulation software developed by Birmingham University is used to validate the timetable produced by the algorithm. The algorithm generated a timetable which was then entered in BRaVE to see if the software could execute the timetable without any infeasibilities arising either due to train collisions or sectional running time violations. The output from BRaVE has shown that the timetable from the algorithm can be executed with minor alterations which arise as a consequence of the difference in the way in which sectional running times are considered in the model as opposed to BRaVE. The optimisation algorithm was consequently deemed to generate feasible timetables.

Chapter 5 presents and analyses the results obtained from the experiments. The chapter starts by introducing the Brighton Main Line, a subsection of which will be used for the experiments. The subsection to be used covers the railway network from Gatwick Airport to Brighton and the time interval to be examined is the morning peak hours between 08:00 and 10:00. The passenger mix during the given time interval is taken to be comprised of 60% commuting, 20% business and 20% leisure passengers. Prior to the initiation of the experiments, the optimisation algorithm was run three times to determine how quickly it converges and it was decided to terminate the algorithm after 200 iterations.

The first series of experiments was aimed to identify whether trains can be sequenced (without scheduling additional trains) in such a way such that the cost of the timetable can be reduced. Experiments were carried out for three demand levels (i.e. low, average and high demand) to determine whether demand levels

can impact the optimised sequence. Results have shown that resequencing can lower the cost of punctuality by evenly distributing the trains which exit the network early. Such an action creates an artificial buffer between the train before and after the train exiting early and this buffer absorbs delays, leading to lower punctuality costs. Demand levels did not appear to have any impact on the optimised sequence.

Analysing the effects of scheduling additional trains was the series of experiments to be carried out. The trains added were taken to be the service from Gatwick Airport and back since that is the route with the highest demand meaning that the most major improvements can be captured by scheduling additional trains serving that route. Three demand levels were examined which were the same as the ones used for the experiments above. Results have shown that, as expected, the optimised number of trains to be scheduled depends on the demand levels. If demand is low, scheduling extra trains will only increase the cost of Punctuality and Journey Time with minimal improvements in the cost of Waiting Time. This is because if train loading falls below a threshold level, no penalty for overcrowding is imposed, leading to the scheduling of additional trains to have an adverse effect on the total cost of the timetable. When demand is sufficiently high, inserting more trains in the timetable reduces the cost of crowdedness but also increases the cost of punctuality. Due to the fact that crowdedness gains diminish while punctuality costs increase exponentially, the optimised number of trains to be scheduled relies on how crowded the train services are.

The last series of experiments in Chapter 5 constructs the Pareto Frontiers for three cost function combinations

- Crowdedness against Punctuality (frontier constructed for average and high demand levels)
- Crowdedness against Journey Time (frontier constructed for average and high demand levels)
- Journey Time against Punctuality (frontier constructed for low, average and high demand levels)

All seven frontiers are expressed with a logarithmic curve since, not only does it fit the data well, it is also meaningful from a conceptual point of view. The Pareto Frontier represents the set of efficient solutions when the cost functions are optimised with different scalar values. This means that depending on which objective is prioritised, the optimal solution can be determined by referring to the appropriate co-ordinates on the Pareto Frontier.

Moreover, when crowdedness is plotted on the x-axis, the slope of the Pareto Frontier reduces as demand levels increase. This is attributed to the fact that, for high demand levels, the existence of standing passengers leads to bigger decreases in the cost of Crowdedness, smoothing the slope of the curve.

6.2 Contribution to the research field

A significant contribution of the research has been the formulation of cost functions to evaluate the non-monetary cost of a railway timetable. As evident from Chapter 2, the vast majority of the literature evaluates a timetable in terms of its network capacity, punctuality and journey time while Waiting time is only rarely measured. This project provides formulations to evaluate punctuality and journey time but in addition it formulates a cost function which calculates the cost of waiting time which is timetable depended. In addition, a cost function is presented which calculates a timetable's crowdedness cost. In the literature, crowdedness cost is only used by the Office of Rail and Road to estimate the impact of investment decisions but, as it is only used on a strategic level, no formulation exists which can be used to evaluate railway timetables.

One of the experiments in Chapter 5 had the purpose of examining how different passenger demand levels can affect the cost of the timetable. To the best of the author's knowledge, no effort has been made by previous authors to examine such a timetabling parameter. This may be attributed to the fact that the decision to carry out the specific experiments can be attributed to the decision to include the Crowdedness cost function. Since the aforementioned cost function has not been considered by any authors, the need to examine different demand levels never arose. The experiments have indeed shown that the demand levels can influence the optimal decision by determining the number of trains to be scheduled. This is important since, until the time of writing, authors have examined the trade-off

between network capacity and punctuality and established their relationship but, without the Crowdedness cost function, they are unable to provide any meaningful method for determining how many trains to schedule given the relationship between capacity and punctuality. For example, it is known that as more trains are added, the cost of punctuality increases exponentially but this does not provide enough information to make a decision on how many trains to schedule. The inclusion of the crowdedness cost function suggests that at very low demand levels no further trains need to be added while as demand levels increase more trains need to be included in the timetable since the reduction in the cost of crowdedness can offset the increases in the cost of punctuality.

In literature, numerous authors attempt to formulate the relationship between network capacity and punctuality. However, when it comes to the rest of the objectives little attempt is being made to understand their relationship. This may be attribute to the fact that the main focus of the authors is the development of new algorithms with the objective functions only serving the purpose of being an input to the algorithm. This has been one of the major targets of this research which has presented the Pareto Frontiers for three different cost function combinations and each combination was further analysed for different demand levels.

Finally, the optimisation procedure developed has not been based on any previous algorithms developed for solving the railway timetabling problem. The need to create timetables from scratch rather than rescheduling and the need to vary the number of trains scheduled created the need to develop something new to tackle the

demands set by this project. However, since the development of an optimisation algorithm was initially out of the scope of this project, there are multiple areas the algorithm can be improved on. A list of potential improvements is given in Section 6.3.

6.3 Future work

A obvious limitation of the project, is the absence of cost functions which estimate the timetable's monetary cost such as the money Network Rail receives for scheduling additional trains and the operational cost of running a service. This was something that was impossible to do due to the inability to access the data to accurately calculate these costs due to data privacy issues which could not be overcome. Consequently, future research can seek to obtain the necessary data will allow for the formulation of these cost functions. Combining the non-monetary cost functions from this research with the monetary cost functions will enable a holistic calculation of the total cost of a railway timetable. This will enable a sensitivity analysis to be carried out to understand the dynamics which govern all the cost functions relevant in the optimisation of railway timetables.

The monetary coefficients used in the research are based on the values proposed by the British Department for Transport. In other countries, different values may be used which will, inevitably, have an impact on the results. This is more obvious from the formulation of the Crowdedness cost function which, as seen from the

experiments, can have a big impact on timetabling decisions. In particular, if significantly different crowdedness multipliers are used, they have the potential to dramatically impact the interaction between the cost functions. Applying this set of cost functions in a country with different travelling time valuations, different results may be reported.

For the purpose of this project, the cost functions were combined using the weighted sum multi-objective optimisation technique (i.e. linear combination of cost functions). However, the problem can be formulated using different techniques such as

Lexicographic optimisation optimises the problem using one objective, then constraints its value and optimises the second objective with the additional constraint imposed.

Goal Programming optimises a single objective function and imposes a soft constraint on the rest of the functions. If any of the objective function constraints is violated, a penalty is imposed which increases according to the value by which the constraint was exceeded.

Data Envelopment Analysis for any feasible solution to the optimisation problem, DEA calculates a score in the range 0 – 1 for each objective function scoring how efficient the objective is for the given solution.

Each of the above techniques can be used depending on the problem requirements (e.g. if one objective is infinitely more important than the rest, lexicographic

optimisation can be used). In particular, DEA can be implemented to efficiency score of each objective when they are optimised under different circumstances (e.g. different crowdedness levels, different monetary coefficients for each objective etc.).

The purpose of the project was the formulation and analysis of cost functions rather than the development of an efficient and effective optimisation procedure. Even though better algorithms have been developed to solve the train timetabling problem, the algorithm developed in this project serves a slightly different purpose and, as it presents the opportunity for further usage, it could be further refined. Future researchers may focus on improving certain aspects of the algorithm to increase its ability to construct efficient timetables. Some of the areas to be improved are

- Modify the way train priorities are determined in cases of conflicts. At the moment, when two trains clash at a junction, priority is given to the train which appears higher in the sequence outputted by the Genetic Algorithm.
- The Hill-Climbing heuristic only adds a specific service (e.g. from Brighton to Gatwick without any intermediary stops) without having the flexibility to schedule different services. It will be interesting to add an additional feature which, in each iteration, will examine the different alternatives and schedule the service which offers the highest cost reductions.
- When the Hill-Climbing heuristic schedules an additional train, the train is placed last in the sequence list. This implies that the train will be scheduled subject to the constraints imposed by all the previous trains scheduled before

it. Therefore, a procedure can be developed which will examine whether it is more beneficial to insert the train in a different place in the sequence list other than the last one.

Even if the changes above are not implemented, it will be worth coding the optimisation algorithm from scratch. When the algorithm was initially developed, the foundations were laid to create something vastly different but, according to the project's changing demands, the algorithm ended up being build in a way that its computational speed suffers significantly. Coding the whole algorithm from scratch now that the procedure has been finalised will help to speed the algorithm up considerably. One benefit of this is the ability to run more Monte-Carlo simulations without the need to devote countless hours in computational time. When these changes have been made, the effectiveness and efficiency of the algorithm can be benchmarked against other optimisation algorithms used for railway timetabling.

Examining the Pareto Frontiers has shown that, when Punctuality was on of the cost functions being analysed, the scatter plot was exhibiting high variability. This is somewhat expected due to the fact that delays are generated randomly but, in some cases (e.g. Figures 5.21 and Figure 5.22) certain data points were lying very far from the Pareto Frontier. The variability can somewhat be reduced by running an increasing number of Monte-Carlo simulations to construct the stochastic timetable. The way the optimisation algorithm was structured and implemented was rendering it impractical to run additional simulations due to it being very time consuming to do so. Future research can therefore carry out the

experiments using a higher number of simulation runs to construct the Pareto Frontiers in order to examine whether the results will change.

The case study presented, only considered passenger trains. However, it will be interesting to examine the impact of scheduling freight trains as well and also whether any monetary coefficients can be assigned to the different activities regarding freight trains (e.g. journey time for freight trains).

Finally, a number of projects are currently under way by Network Rail (e.g. DE-DOTS) which aim to improve the robustness and dependability of operations while also making better use of railway capacity and minimising energy consumption [21]. These project though focus on the algorithmic aspects of timetable optimisation so the work presented in this project could supplement such project by providing a set of cost functions which can be used to develop optimised train timetables. This can provide an opportunity for implementation in the railway industry since all the tools developed by academia supplement each other and contribute towards the vision that Network Rail has outlined for the future of traffic management systems.

Appendix A

Origin-destination matrix for the ECML

TABLE A.1: Origin-destination matrix between Alexandra Palace and Hatfield

		Destination									
		Alexandra Palace	Bowes Park	New Southgate	Oakleigh Park	New Barnet	Hadley Wood	Potters Bar	Brookmans Park	Welham Green	Hatfield
Origin	Alexandra Palace	0.000	0.100	0.067	0.067	0.067	0.067	0.100	0.067	0.067	0.400
	Bowes Park	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	New Southgate	1.000	0.000	0.000	0.060	0.060	0.060	0.200	0.060	0.060	0.500
	Oakleigh Park	0.900	0.000	0.100	0.000	0.060	0.050	0.200	0.050	0.050	0.600
	New Barnet	0.800	0.000	0.100	0.100	0.000	0.067	0.200	0.067	0.067	0.600
	Hadley Wood	0.700	0.000	0.100	0.100	0.100	0.000	0.100	0.100	0.100	0.700
	Potters Bar	0.600	0.000	0.100	0.100	0.100	0.100	0.000	0.100	0.100	0.800
	Brookmans Park	0.600	0.000	0.075	0.075	0.075	0.075	0.100	0.000	0.100	0.900
	Welham Green	0.500	0.000	0.075	0.075	0.075	0.060	0.200	0.060	0.000	1.000
Hatfield	0.500	0.000	0.075	0.075	0.075	0.050	0.200	0.050	0.050	0.000	

Appendix B

Timetables generated for validation

TABLE B.1: Arrival times as generated by the optimisation procedure

Service	Alexandra Palace	Bowes Park	New Southgate	Oakleigh Park	New Barnet	Hadley Wood	Potters Bar	Brookmans Park	Welham Green	Hatfield
S62	08:00:00		08:03:00	08:06:30	08:08:30	08:11:00	08:15:00	08:18:00	08:20:00	08:23:30
S857	08:03:00		08:06:00	08:09:30	08:11:30	08:14:00	08:18:00	08:21:00	08:23:00	08:26:40
S31	08:29:05		08:26:05	08:22:35	08:20:20	08:17:20	08:12:50	08:09:30	08:06:40	08:03:00
S201	08:06:00	08:08:40								
S32	08:32:05		08:29:05	08:25:35	08:23:20	08:20:20	08:15:50	08:12:30	08:09:40	08:06:00
S1031	08:15:00		08:14:04	08:12:49	08:12:19	08:11:23	08:10:00	08:09:04	08:08:30	08:06:00
S33	08:35:05		08:32:05	08:28:35	08:26:20	08:23:20	08:18:50	08:15:30	08:12:40	08:09:00

TABLE B.2: Arrival times as generated by BRaVE

Service	Alexandra Palace	Bowes Park	New Southgate	Oakleigh Park	New Barnet	Hadley Wood	Potters Bar	Brookmans Park	Welham Green	Hatfield
S62	08:00:00		08:03:00	08:06:37	08:08:35	08:11:17	08:15:17	08:18:11	08:20:15	08:23:31
S857	08:03:00		08:05:51	08:09:18	08:11:20	08:14:11	08:18:19	08:21:25	08:23:16	08:26:47
S31	08:28:54		08:25:54	08:22:28	08:20:11	08:17:07	08:12:41	08:09:08	08:06:34	08:03:00
S201	08:06:00	08:08:29								
S32	08:31:54		08:28:54	08:25:28	08:23:11	08:20:20	08:15:03	08:12:18	08:09:34	08:06:00
S1031	08:15:00		08:14:04	08:12:49	08:12:19	08:11:23	08:10:00	08:09:04	08:08:30	08:06:00
S33	08:35:21		08:32:27	08:28:38	08:26:30	08:23:29	08:19:02	08:15:18	08:12:34	08:09:00

Appendix C

Origin-destination matrix for the BML

TABLE C.1: Origin-destination matrix between Gatwick Airport and Brighton

	Gatwick Airport	Three Bridges	Balcombe	Haywards Heath	Wivelsfield	Burgess Hill	Hassocks	Preston Park	Brighton
Gatwick Airport	0	0.250	0.050	0.050	0.050	0.050	0.50	0.050	0.450
Three Bridges	1	0	0.083	0.083	0.083	0.083	0.083	0.083	0.500
Balcombe	0.800	0.200	0	0.100	0.100	0.100	0.100	0.100	0.500
Haywards Heath	0.500	0.300	0.200	0	0.125	0.125	0.125	0.125	0.500
Wivelsfield	0.500	0.300	0.100	0.100	0.000	0.167	0.167	0.167	0.067
Burgess Hill	0.500	0.300	0.067	0.067	0.067	0.000	0.250	0.250	0.500
Hassocks	0.500	0.300	0.050	0.050	0.050	0.050	0.000	0.200	0.500
Preston Park	0.500	0.300	0.040	0.040	0.040	0.040	0.040	0.000	1.000
Brighton	0.500	0.300	0.033	0.033	0.033	0.033	0.033	0.033	0.000

Appendix D

Terminology

Allowance time The time added into the nominal timetable to compensate the additional train sectional running times, dwell times and other scheduled process times due to unavoidable variability of physical characteristics, driver behaviours, passengers boarding and alighting variations and other potential influencing factors to train operations in real life conditions. They are included by increasing the scheduled SRTs of trains.

Arrival delay A deviation of the arrival time from the scheduled arrival time at a station.

Block signal A stop signal that controls the entrance to or signifies the termination of a block or signal section and any other stop signal within station limits.

Blocking time The time interval in that a section of track is allocated to the exclusive use of one train and therefore blocked to other trains.

Buffer time The time added into the nominal timetable (between train slots) to reduce or avoid propagation of knock-on delays among running trains due to initial and/or primary train delays.

Corridor All possible journey routes (main route or alternative routes), according to market needs, between a defined source and target.

Crossing An assembly of rails that enables two tracks or two pair of tracks to cross each other at grade.

Delay The deviation from either a scheduled event or process time of this train.

Departure delay A deviation of the departure time from the scheduled arrival time at a station.

Dwell time The elapsed time from the time that a train stops at a station platform until it starts moving again.

Flat junction Junctions which lead to conflicting moves between trains going in one direction and trains coming in the opposite direction (flying junction is the opposite).

Flighting Running consecutive trains of a similar type. This minimises the space used by each group of trains and is used through the Channel Tunnel.

Freight operating company A company with access rights to operate freight trains on the railway network.

Headway The necessary time interval or space between two successive trains on the same track.

Infrastructure manager A body responsible for development, operation and maintenance of the railway infrastructure (Network Rail is the main IM for the mainline network in the UK).

Infrastructure The fixed and capital equipment needed for running, maintaining, signalling and dispatching trains.

Knock-on delay The delay cause to a train as a result of a delay to another train.

Line A link between two large nodes and usually the sum of more than one line section.

Line sections The part of a line, in which the traffic mix and the number of trains as well as the infrastructure and signalling conditions do not change fundamentally.

Network capacity The number of trains that can operate in a rail network in a given time period, reflecting factors such as junction interactions, terminal capabilities, the mix of train speeds and the number and order of trains of different speed capabilities and stopping patterns called for by commercial and regulatory requirements.

Node Points of a network in which at least two lines converge (can be stations or junctions).

Overlap The distance beyond a stop signal up to which the line must be clear before the previous signal can show a proceed aspect.

Passenger journey The combination between the place of embarkment and the place of disembarkment of the passengers conveyed by rail whichever itinerary is followed.

Primary delay A delay generated within the network and not caused by other trains.

Punctuality Defined by Network Rail as the percentage of the trains that arrive at a location with a delay not exceeding the allowance time.

Public Performance Measure (PPM) The national standard for measuring punctuality is the percentage of trains that arrive at their final destination within ten minutes of the advertised time.

Route Consecutive lines and nodes as a whole, between a defined source and target.

Railway network A train system or a particular area including all train running elements which can communicate with other networks.

Siding The term siding may refer to any track where railway vehicles may be left (i.e. are not an operating train for the time being). The duration that such vehicles are in a siding may vary from few minutes to years.

Track circuit A portion of railway line having fixed boundaries and providing information on its state of occupancy to the signalling system. Within this

standard, this traditional name does not preclude alternative forms of train detection.

Train operating company A company with access rights to operate passenger trains on the railway network.

Signal section The line between two stop signals, whether or not these are within the control of the same signal box.

Skip stop patterns Using pairs or patterns of trains to cover all stations using semi-fast services with different stopping patterns. This avoids running slow all-stations services which use more capacity.

System capacity The total capacity of the railway system to carry passengers or freight. This is the resultant of passenger capacity of each vehicle of payload of each freight wagon, the number of vehicles on each train and the Network capacity (see above).

Timetabling The process for constructing a schedule outlining the arrival and departure time of all the services run from all the stations in their path. The schedule must adhere to a list of operational constraint (e.g. minimum headway requirements).

Train loading The number of passengers on board relative to the train's seating capacity.

Train path That part of capacity of the railway infrastructure which is necessary to schedule or run a train with a requested speed profile.

References

- [1] Abril, M., Barber, F., Ingolotti, L. P., Salido, M. A., Tormos, P., and Lova, A. L. 2008. An assessment of railway capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(1), pp. 774–806.
- [2] Albrecht, A., Howlett, P., Pudney, P., Vu, X., and Zhou, P. 2015. The key principles of optimal train control - Part 1: Formulation of the model, strategies of optimal type, evolutionary lines, location of optimal switching points. *Transportation Research Part B: Methodological*, 1(1), pp. 1–27.
- [3] Albrecht, A., Howlett, P., Pudney, P., Vu, X., and Zhou, P. 2015. The key principles of optimal train control - Part 2: Existence of an optimal strategy, the local energy minimization principle, uniqueness, computational techniques. *Transportation Research Part B: Methodological*, 1(1), pp. 1–30.
- [4] Albrecht, T. 2009. Automated timetable design for demand-oriented service on suburban railways. *Public Transport*, 1(1), pp. 5–20.

-
- [5] Barber, F. and Ingolotti, L. and Lova, A. and Tormos, P. and Salido, M., . 2009. Meta-heuristics and Constraint-Based Approaches for Single-Line Railway Timetabling. *Robust and Online Large-Scale Optimization: Models and Techniques for Transport Systems*. pp. 145–181. Springer Berlin Heidelberg.
- [6] Beagles, A. and Fletcher, D. 2013. The aerodynamics of freight; approaches to save fuel by optimising the utilisation of container trains. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(6), pp. 635–643.
- [7] Ben-Tal, A. and Nemirovski, A. 1999. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1), pp. 1–13.
- [8] Bertsimas, D., Brown, D. B., and Caramanis, C. Massachusetts Institute of Technology. 2007. Theory and applications of robust optimisation. URL <http://web.mit.edu/dbertsim/www/papers/Robust%20Optimization/Theory%20and%20applications%20of%20robust%20optimization.pdf>.
- [9] Buchanan, C. and Volterra Consulting, . Volterra. 2007. The economic benefits of crossrail. URL <http://volterra.co.uk/wp-content/uploads/2013/02/Economic-Benefits-of-Crossrail.pdf>.
- [10] Burdett, R. L. and Kozan, E. 2006. Techniques for absolute capacity determination in railways. *Transportation Research Part B: Methodological*, 40(8), pp. 616–632.

-
- [11] Bussieck, M. R., Winter, T., and Zimmermann, U. T. 2009. Discrete optimization in public rail transport. *Mathematical Programming* 79, 79(1-3), pp. 415–444.
- [12] Butcher, L., . House of Commons. 2015. Railways: Passenger franchises. URL researchbriefings.files.parliament.uk/documents/SN01343/SN01343.pdf.
- [13] Cai, X. and Goh, C. J. 1994. A fast heuristic for the train scheduling problem. *Computers and Operations Research*, 21(5), pp. 499–510.
- [14] Caprara, A., Fischetti, M., and Toth, P. 2002. Modeling and solving the train timetabling problem. *Operations Research*, 50(5), pp. 851–861.
- [15] Carey, M. 1994. Reliability of interconnected scheduled services. *European Journal of Operational Research*, 79(1), pp. 51–72.
- [16] Carey, M. 1999. Ex ante heuristic measures of scheduled reliability. *Transportation Research Part B: Methodological*, 33(7), pp. 473–494.
- [17] Carey, M. and Carville, S. 2003. Scheduling and platforming trains at busy complex stations. *Transportation Research Part A: Policy and Practice*, 37(3), pp. 195–224.
- [18] Carey, M. and Kwiecinsky, A. 1995. Properties of expected costs and performance measures in stochastic models of scheduled transport. *European Journal of Operational Research*, 83(1), pp. 182–199.

-
- [19] Carlier, J. and Pinson, E. 2004. Jackson's pseudo-preemptive schedule and cumulative scheduling problems. *Discrete Applied Mathematics*, 1(30), pp. 80–94.
- [20] Chen, L. and Roberts, C. RSSB. 2012. Defining the Optimisation Function for Timetabling as Part of Future Traffic Regulation (FuTRO). URL www.sparkrai.org.
- [21] Chow, A., Heydecker, B., Fujiyama, T., and Xu, F. DEDOTS: Developing and Evaluating Dynamic Optimisation for Train Control Systems. URL <https://www.ucl.ac.uk/resilience-research/research/dedots>.
- [22] Cordeau, J. F., Toth, P., and Vigo, D. 1998. A survey of optimization models for train routing and scheduling. *Transportation Science*, 32(4), pp. 380–404.
- [23] Corman, F., D'Ariano, A., Hansen, I., and Pacciarelli, D. 2011. Optimal multi-class rescheduling of railway traffic. *Journal of Rail Transport Planning & Management*, 1(1), pp. 14–24.
- [24] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. 2013. *Introduction to Algorithms*. PHI Learning, Cambridge.
- [25] Daganzo, C. F. 2007. *Fundamentals of Transportation and Traffic Operations*. Pergamon, New York.
- [26] D'Ariano, A. 2008. Improving real-time train dispatching: Models, algorithms and applications. Doctoral Dissertation, TU Delft: Germany.

-
- [27] D'Ariano, A., Pacciarelli, D., and Pranzo, M. 2007. A branch and bound algorithm for scheduling trains in a railway network. *European Journal of Operational Research*, 183(2), pp. 643–657.
- [28] De Kort, A. F., Heidergott, B., and Ayhan, H. 2003. A probabilistic (max, +) approach for determining railway infrastructure capacity. *European Journal of Operational Research*, 148(1), pp. 644–661.
- [29] Department for Transport, . Department for Transport. 2012. The High Level Output Specification 2012: Railways Act 2005 statement. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/3641/railways-act-2005.pdf.
- [30] Department for Transport, . Department for Transport. 2014. Brighton Main Line: Emerging Capacity Strategy for CP6. Pre-Route Study report for DfT. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/306997/brighton-main-line-interim-report.pdf.
- [31] Dingler, M. H., Lai, Y., and Barkan, C. P. L. 2009. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of Transportation Research Board*, No. 2117. pp. 41–49. Transportation Research Board of the National Academies, Washington, D.C.
- [32] Dorfman, M. J. and Medanic, J. 2004. Scheduling trains on a railway network using a discrete event model of railway traffic. *Transportation Research Part B: Methodological*, 38(1), pp. 81–98.

- [33] Ehrgott, M. 2000. *Multicriteria Optimization*. Springer Berlin Heidelberg, Berlin, 2nd edition.
- [34] Elkin, S., . Network Rail. 2016. Timetable Planning Rules: Sussex. URL <http://www.networkrail.co.uk/browse%20documents/Rules%20of%20The%20Route/Viewable%20copy/TPRyearYY/ktYYp.pdf>.
- [35] Finnigan, L. Rail punctuality drops to worst level in a decade, as new figures show top 10 worst performing lines. URL <http://www.telegraph.co.uk/news/2016/05/12/rail-network-punctuality-drops-to-worst-level-in-almost-a-decade/>.
- [36] Fischetti, M. and Monaci, M., . 2009. Light Robustness. *Robust and Online Large-Scale Optimization: Models and Techniques for Transport Systems*. pp. 61–84. Springer Berlin Heidelberg.
- [37] Ghoseiri, K., Szidarovszky, F., and Asgharpour, M. J. 2004. A multi-objective train scheduling problem and solution. *Transportation Research Part B: Methodological*, 38(10), pp. 927–952.
- [38] Gibson, S., Cooper, G., and Ball, B. 2002. The evolution of capacity changes on the UK rail network. *Journal of Transport Economics and Policy*, 36(2), pp. 341–354.
- [39] Goossens, J. W., Hoesel, S.van , and Kroon, L. 2004. A branch-and-cut approach for solving railway line-planning problems. *Transportation Science*, 38(3), pp. 379–393.

- [40] Goverde, R., Corman, F., and D’Ariano, A. 2013. Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions. *Journal of Rail Transport Planning & Management*, 3(3), pp. 78–94.
- [41] Hallowell, S. F. and Harker, P. T. 1998. Predicting on-time performance in scheduled railroad operations: Methodology and application to train scheduling. *Transportation Research Part A: Policy and Practice*, 32(4), pp. 279–295.
- [42] Higgins, A., Kozan, E., and Ferreira, L. A. 1996. Optimal scheduling of trains on a single line track. *Transportation Research Part B: Methodological*, 30(2), pp. 147–161.
- [43] High Speed Two (HS2) Limited, . Department for Transport. 2013. The economic case for HS2. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/365065/S_A_1_Economic_case_0.pdf.
- [44] Institution of Railway Operators, . 2014. *Operators Handbook*. Institution of Railway Operators, Stafford, United Kingdom, 2nd edition.
- [45] Klages, S. G. 2010. Algorithmic railway capacity allocation in a competitive european railway market. Doctoral Dissertation, RWTH Aachen University: Germany.
- [46] Kraay, D. R. and Harker, P. T. 1994. Real-time scheduling of freight railroads. *Transportation Research Part B: Methodological*, 29B(3), pp. 213–229.

- [47] Kroon, L., Maroti, G., Helmrich, M. R., Vromans, M., and Dekker, R. 2008. Stochastic improvement of cyclic railway timetables. *Transportation Research Part B: Methodological*, 42(6), pp. 553–570.
- [48] Li, F., Gao, Z., Li, K., and Yang, L. 2008. Efficient scheduling of railway traffic based on global information of train. *Transportation Research Part B: Methodological*, 42(10), pp. 1008–1030.
- [49] Liebchen, C., Schachtebeck, M., Schobel, A., Stiller, S., and Prigge, A. 2010. Computing delay resistant railway timetables. *Computers and Operational Research*, 37(5), pp. 857–868.
- [50] Lo, H. K. and Chow, A. H. F. 2004. Control strategies for over-saturated traffic. *ASCE Journal of Transportation Engineering*, 130(4), pp. 466–478.
- [51] Mackie, P. J., Jara-Diaz, S., and Fowkes, A. S. 2001. The value of travel time saving in evaluation. *Transportation Research Part E: Logistics and Transportation Review*, 37(2-3), pp. 91–106.
- [52] Mackie, P. J., Wardman, M., Fowkes, A. S., Whelan, G., Nellthorp, J., and Bates, J. University of Leeds. 2003. Values of travel time savings in the UK. URL http://eprints.whiterose.ac.uk/2079/2/Value_of_travel_time_savings_in_the_UK_protected.pdf.
- [53] McNulty, R., . Department for Transport and Office of Rail and Road. 2011. Realising the Potential of GB Rail: Report of the Rail Value for Money Study. URL http://orr.gov.uk/_data/assets/pdf_file/0009/1710/rail-vfm-summary-report-may11.pdf.

- [54] Meester, L. E. and Muns, S. 2007. Stochastic delay propagation in railway networks and phase-type distributions. *Transportation Research Part B: Methodological*, 41(2), pp. 218–230.
- [55] Mott MacDonald Limited, . Department for Transport. 2009. Productive Use of Rail Travel Time and the Valuation of Travel Time Savings for Rail Business Travellers. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/4003/productive-use-of-travel-time.pdf.
- [56] Mussone, L. and Calvo, R. W. 2013. An analytical approach to calculate the capacity of a railway system. *European Journal of Operational Research*, 228(1), pp. 11–23.
- [57] Nachtigall, K. 1996. Periodic network optimization with different arc frequencies. *Discrete Applied Mathematics*, 69(1-2), pp. 1–17.
- [58] Nachtigall, K. and Voget, S. 1996. A genetic algorithm approach to periodic railway synchronization. *Computers and Operations Research*, 23(5), pp. 453–463.
- [59] Network Rail, . Methodology: Public Performance Measure (PPM). URL <https://dataportal.orr.gov.uk/displayreport/report/html/ea0ad5a2-daca-47ec-a72e-b6e50a36a3f1>.
- [60] Network Rail, . Performance. URL <http://www.networkrail.co.uk/about/performance/>.

- [61] Network Rail, . Station usage 2014-15 data. URL <http://orr.gov.uk/statistics/published-stats/station-usage-estimates>.
- [62] Network Rail, . Network Rail. 2010. Route Plans 2010: Route Plan G East Coast & North East. URL <http://www.networkrail.co.uk/RoutePlans/PDF/RouteG-EastCoastandNorthEast.pdf>.
- [63] Network Rail, . Network Rail. 2010. East Coast Main Line 2016 Capacity Review: An addendum to the East Coast Main Line Route Utilisation Strategy. URL <http://www.networkrail.co.uk/browse%20documents/rus%20documents/route%20utilisation%20strategies/east%20coast%20main%20line/east%20coast%20main%20line%202016%20capacity%20review/east%20coast%20main%20line%202016%20capacity%20review.pdf>.
- [64] Network Rail, . Department for Transport. 2015. Our History. URL <http://www.networkrail.co.uk/aspx/729.aspx>.
- [65] Network Rail, . Department for Transport. 2015. Network Statement 2016. URL <http://www.networkrail.co.uk/aspx/3645.aspx>.
- [66] Network Rail, . Department for Transport. 2015. Schedule 8: Performance Regime. URL http://www.networkrail.co.uk/browse%20documents/track%20access/2%20completed%20consultations/2008/2008.12.19%20nx%20east%20anglia%20new%20track%20access%20contract%20-%20consultation%20closed%2015%20january%202009/1er%20new%20tac%20schedule%208%20%285248731_5%29%20redacted.pdf.

- [67] Network Rail, . Network Rail. 2015. Book YA working Timetable: Passenger train services. URL [http://www.networkrail.co.uk/browse%20documents/timetables/working%20timetable%20\(wtt\)/2%20-%20december%202015%20-%20may%202016/ya/ya01.pdf](http://www.networkrail.co.uk/browse%20documents/timetables/working%20timetable%20(wtt)/2%20-%20december%202015%20-%20may%202016/ya/ya01.pdf).
- [68] Odijk, M. 1996. A constraint generation algorithm for the construction of periodic railway timetables. *Transportation Research Part B: Methodological*, 30(6), pp. 455–464.
- [69] Office of Rail and Road, . ORR. 2015. Methodology: Public Performance Measure (PPM). URL http://orr.gov.uk/__data/assets/pdf_file/0015/4425/performance-quality-report.pdf.
- [70] Osuna, E. E. and Newell, G. F. 1972. Control strategies for an idealized public transportation system. *Transportation Science*, 6(1), pp. 52–72.
- [71] Papadimitriou, C. and Steiglitz, K. 2000. *Combinatorial Optimization: Algorithms and Complexity*. Dover Books on Computer Science, New York.
- [72] Parry-Jones, R., . Network Rail. 2012. Technical Strategy. URL <http://www.networkrail.co.uk/publications/technical-strategy/>.
- [73] Pavlides, A., Chow, A. H. F., and Baker, C. 2015. A study of cost functions for timetabling national railway operations. *Proceedings of the 47th Annual Conference of Universities Transport Study Group*, January 5-7, London, UK.
- [74] Peterson, A. 2012. Towards a robust traffic timetable for the Swedish Southern Mainline. In Brebbia, C. A., Tomii, N., and Tzieropoulos, P., editors,

- Computers in Railways XIII: Computer System Design and Operation in the Railway and Other Transit Systems*, September, pp. 473–484.
- [75] Rail Executive, . Department for Transport. 2015. Rail franchise schedule. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/363173/oct-2014-rail-franchise-schedule.pdf.
- [76] Reeves, C., . 2003. Genetic Algorithms. *Handbook of Metaheuristics*. pp. 109–139. Kluwer Academic Publishers.
- [77] Roberts, C., Chen, L., Lu, M., Dai, L., Bouch, C., Nicholson, G., and Schmid, F. 2013. *A framework for developing an objective function for evaluating work package solutions (Cost function) - Project report: Optimal Networks for Train Integration Management across Europe (ONTIME)*. European Commission Seventh Framework Programme (FP7).
- [78] Rowlatt, A., . Department for Transport. 2015. Understanding and Valuing the Impacts of Transport Investment: Values of travel time savings. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470998/Understanding_and_Valuing_Impacts_of_Transport_Investment.pdf.
- [79] Russell, S. and Novig, P. 2010. *Artificial Intelligence: A Modern Approach*. Pearson, New Jersey.

-
- [80] Sama, M., Meloni, C., D'Ariano, A., and Corman, F. 2015. A multi-criteria decision support methodology for real-time train scheduling. *Journal of Rail Transport Planning & Management*, 5(3), pp. 146–162.
- [81] Sameni, M. K. and Preston, J. M. 2012. Value for railway capacity: Assessing efficiency of operators in Great Britain. *Transportation Research Record: Journal of Transportation Research Board*, No. 2289. pp. 134–144. Transportation Research Board of the National Academies, Washington, D.C.
- [82] Schwanhausser, W. 1994. The status of German railway operations management in research and practise. *Transportation Research Part A: Policy and Practice*, 28(6), pp. 495–500.
- [83] Suteewong, W. 2006. Algorithms for solving the train dispatching problem for general networks. Doctoral Dissertation, University of Southern California: USA.
- [84] Tirachini, A., Hensher, D. A. and Rose, J. M., . 2013. Crowding in public transport systems: Effects on users, operations and implications for the estimation of demand. *Transportation Research Part A: Policy and Practice*, 53(1), pp. 36–52.
- [85] Transport appraisal and strategic modelling division, . Department for Transport. 2009. TAG UNIT 3.5.6: Values of time and vehicle operating costs. URL www.persona.uk.com/a5dunstable/deposit-docs/DD-096.pdf.

- [86] Transport appraisal and strategic modelling division, . Department for Transport. 2010. New Lines Programme: Demand forecasting technical note. URL www.networkrail.co.uk/5879_Demandforecastingtechnicalnote.pdf.
- [87] Union Internationale des Chemins de fer, . Union Internationale des Chemins de fer. 2004. UIC Code 406: Capacity. URL <http://banportalen.banverket.se/Banportalen/upload/1753/HandbokUIC406.pdf>.
- [88] Vansteenwegen, P. and Van Oudheusden, D. 2006. Developing railway timetables which guarantee better service. *European Journal of Operational Research*, 173(1), pp. 337–350.
- [89] Vansteenwegen, P. and Oudheusden, D.van . 2007. Decreasing the passenger waiting time for an intercity rail network. *Transportation Research Part B: Methodological*, 41(4), pp. 478–492.
- [90] Wardman, M. 1998. The value of travel time - A review of British evidence. *Journal of Transport Economics and Policy*, 32(3), pp. 285–316.
- [91] Wardman, M. 2004. Public transport values of time. *Transport Policy*, 11(1), pp. 363–377.
- [92] Wardman, M. and Whelan, G. 2010. Twenty years of rail crowding: Evidence and lessons from British experience. *Transport Reviews: A Transnational Transdisciplinary Journal*, 31(3), pp. 379–398.
- [93] Wen, T., Lyu, X., Kirkwood, D., Chen, L., Constantinou, C., and Roberts, C. 2015. Co-Simulation Testing of Data Communication System Supporting

-
- CBTC. *IEEE 18th International Conference on Intelligent Transport Systems*, September 15-18, Canary Islands, Spain.
- [94] Xu, F., Heydecker, B., and Chow, A. H. F. 2015. Optimisation framework for rail traffic regulation and speed optimisation at a single junction. *Proceedings of International Conference of RailTokyo*, March 23-26, Tokyo, Japan.
- [95] Yuan, J. 2006. Stochastic modelling of train delays and delay propagation in stations. Doctoral Dissertation, Delft University of Technology: Netherlands.
- [96] Yuan, J. and Hansen, I. 2007. Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B: Methodological*, 41(2), pp. 202–217.