# Understanding the Relationship Between Capacity Utilisation and Performance and the Implications for the Pricing of Congested Rail Networks

John Alexander Haith

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
Institute for Transport Studies

February, 2015

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Elements of the work contained in this thesis have previously appeared in the published paper:-

Haith, J., Johnson, D. and Nash, C. 2014.The Case for Space: the Measurement of Capacity Utilisation, its Relationship with Reactionary Delay and the Calculation of the Capacity Charge for the British Rail Network. *Transportation Planning and Technology* **37** (1) February 2014 Special Issue: Universities' Transport Study Group UK Annual Conference 2013.

Where there is specific use of the contents of the above paper in this thesis reference is made to it in the appropriate part of the text. However, general use of the work contained in the paper is particularly made in Chapter 5 (Methodology), Chapter 6 (The Data Set) and Chapter 7 (Results).

It should also be noted that all research and analysis contained in this thesis (and the paper) was conducted by the candidate. Secondly, substantial additional analysis was conducted between the finalisation of the paper and the writing of this thesis meaning that the results of the research have expanded significantly.

## Acknowledgements

A large number of people have provided practical help or encouragement in the research behind this thesis and its writing. In order to properly thank those who have given assistance it is necessary to explain how I came to be writing a thesis on rail capacity utilisation in the first place.

After graduating from The University of Leeds in 1990 with a B.A.(Hons) in Geography I went to work for a transport consultancy as a planner working on road schemes. They encouraged me to undertake a part time M.Sc (Eng) in Transport Planning and Engineering at the Institute for Transport Studies at Leeds to widen my understanding. It was here that I encountered two people that were instrumental in my subsequently undertaking a Ph.D in a rail related matter at Leeds. Firstly, I got to know a fellow student called Doug Desmond who was working for the company called Railtrack who would go onto become the Infrastructure Owner once Britain’s railways were privatised. Doug would become my first manager when I went to work for Railtrack. At the same time I attended a number of stimulating lectures on rail issues by Professor Chris Nash.

I have much to thank Doug for. Most importantly of all he was instrumental in my meeting my wife-to-be who at the time held a similar job in Glasgow to mine in York.

In the end I worked for Railtrack (and its subsequent reincarnation as Network Rail) for 17 years. Much of this time was in the train planning field where I held a number of senior managerial positions including a spell as a Timetable Manager for a large part of the British rail network. During my time in train planning I received excellent advice and encouragement on the intricacies of creating a timetable from numerous people. The understanding I gained during this time has been instrumental in my developing and using the various capacity utilisation measures described in this thesis. It has also helped me understand and interpret the performance data. Finally it has also helped my comprehension of what is likely to be possible and what is not possible in terms of the recommendations made in this thesis.

I also owe a debt of gratitude to Richard O’Brien who as Operational Planning Manager at Network Rail assisted a move into Strategic Access Planning and introduced me to the concept of The Theory of Constraints. It was under his mentorship that I pursued the idea towards the end of my time

at Network Rail that this concept (that the flow through a system is dictated by the flow though its principle constraint) would be a useful way to investigate rail capacity utilisation. It was his support that gave me the confidence to apply for a Ph.D at Leeds in this field. In deciding to pursue this wish I first approached Professor Nash and it is thanks to his help and that of Daniel Johnson that I was accepted.

Both Chris and Dan have proved excellent supervisors. They have provided considerable encouragement but also questioned my analysis and writing where necessary. They have helped tease out my thinking into fully formed concepts. They have both leant me their considerable weight of experience. It is true to say that this thesis would not exist without their assistance. There are also obviously a large number of other people associated with the Institute for Transport Studies who have provided advice, encouragement and constructive criticism along the way.

During my studies I have also been lucky enough to present a paper at two external conferences. The experience gained in doing so together with the constructive criticism received has helped enormously.

I would also like to thank Network Rail for their provision of the data required to carry out the analysis described in this thesis. I would particularly like to thank Nigel Salmon, my last manager at the company, who kindly provided me with the performance data set after spending some considerable time discussing with me my requirements. Network Rail is also to be thanked for giving me the opportunity to both contribute to the consultation process for the 2013 Recalibration of the Capacity Charge and also to take part in a workshop for the rail industry on developing measures for understanding and optimising capacity utilisation.

Another acknowledgement that I need to make is to my GP and other health professionals in the area. Prior to starting my Ph.D I was diagnosed with a serious illness. Although, at times it has been difficult I have received excellent support from the National Health Service. The University has also been fantastic with regard to the time off I have had to take.

I also wish to acknowledge the support and encouragement of my close friends, the members of my local church and my wider family. Having my parents live ‘just round the corner’ has led to a weekly lunch at their house during my studies which has become a much anticipated and needed break particularly during the writing up phase.

Finally, I wish to acknowledge the love, support and encouragement of my wonderful wife and two amazing daughters. It is to these three people I dedicate this thesis.

John Haith February 2015, Leeds.

# Abstract

There is a growing demand for rail travel in this country which is difficult to satisfy. The result is increased congestion on Britain's railways. One feature of rail infrastructure congestion is a direct link between capacity utilisation and reactionary delay. The latter is the secondary delay that an already late train causes to a following train.

This thesis re-examines the relationship between capacity utilisation and performance (as expressed by the level of reactionary delay). It compares the effectiveness of the standard measure of capacity utilisation in Britain (the Capacity Utilisation Index or CUI) with amongst others a measure developed in the Netherlands (the Heterogeneity measure or HET) which uses a radically different approach. The analysis presented in this thesis finds that HET which measures how capacity is used through the spacing of trains, is a more effective predictor of the levels of reactionary delay than CUI which simply measures how much capacity is used. In both cases though, an exponential relationship between capacity utilisation and reactionary delay is preferred, reinforcing the work of previous researchers.

In 2002 a congestion charge, called The Capacity Charge, was introduced in Britain. The idea was to encourage the Infrastructure Owner (now known as Network Rail) to accommodate more traffic whilst working with train operators to optimise capacity utilisation on the network. The Capacity Charge is based on the relationship between CUI and reactionary delay. However, this thesis shows that HET based tariffs would charge more for congestion than CUI based tariffs. In addition there is a greater differential between peak and off-peak charges. One conclusion is that CUI undercharges for congestion due to its failure to account for the impact of train 'bunching'.

# Table of Contents

## List of Tables

## List of Figures

# Preface

“For my part, I travel not to go anywhere, but to go. I travel for travel’s sake. The great affair is to move” Robert Louis Stevenson (1850-1894). Scottish novelist, poet, essayist and travel writer.

“You and I come by road or rail, but economists travel by infrastructure” Margaret Thatcher (1925-2013) British Prime Minister 1979-1990.

# Chapter 1
# Introduction

## 1.1 Background and Rationale

This thesis is concerned with the relationship between capacity utilisation and performance and the implications of the findings on this for levying a congestion charge on  rail networks with particular reference to Britain. The privatisation of the network in this country in 1994 has had a significant influence on the nature of railway operations and therefore on the answers to the questions posed by this thesis. It is therefore appropriate to first briefly review the nature of the post-privatised rail industry.

The privatisation of the railways in Britain vertically separated the industry. Ownership of the rail infrastructure was given to a newly created company called Railtrack. Following financial difficulties this was replaced by a company called Network Rail. Network Rail is responsible for safely operating and managing everything that ‘does not move’ on the rail network. This includes the planning and controlling of train movements; managing the day-to-day operations at the countries major stations and planning and executing the maintenance of the rail network. As a private sector monopoly owner and operator, Network Rail’s actions are subject to the scrutiny of the Office of Rail Regulation (or ORR). One of the roles of the ORR is to encourage competition on Britain’s rail network and so reduce costs whilst improving the level of choice for customers. The ORR is also responsible for agreeing with Network Rail how much money the latter will receive to operate the rail network. One aspect of this is to encourage the optimum use of the infrastructure.

However, responsibility for the operation of actual train services devolved to a number of other newly created organisations. The vast majority of Britain’s passenger services are operated by Franchise holders. Each of these companies has won the right through a competitive process to operate a set of services specified by the Department for Transport (or in the case of Scotland and Wales by their respective devolved Governments). The successful franchisees either pay the government a fee for operating commercially attractive franchises or instead receive a subsidy for those franchises whose services are being provided from a welfare perspective.  A small number of other passenger operators run services outside the franchise process. These are called Open Access operators and are strictly commercial companies. Their access to the rail network is however still regulated by the ORR.  Freight services are also operated by a number of specialist rail companies. Once again these are operated on a purely commercial basis but this time their access is not subject to regulation. Instead their success depends on winning contracts with customers and obtaining commercially attractive train paths from Network Rail.

There are also a large number of other types of organisation involved in the operation of Britain's rail network. These include those responsible for the purchase and hire of rolling stock and those undertaking specialist engineering work for the renewal of rail infrastructure.

The privatisation of the rail network has created the need for a myriad of detailed legal agreements, with associated incentives and penalties, between the various parties to the agreement. The importance of monitoring and understanding the reliability of railway services has therefore increased significantly, due to the development of these agreements between the various parties as well as the investment of substantial sums of money, both public and private sector, in Britain's rail network. One reflection of the latter is the ORR's interest in the performance of the timetable in Britain and its willingness to fine Network Rail when it fails to meet agreed performance bench-marks. A benefit of the importance of monitoring is that detailed information is available on the amount, location and cause of delay. Detailed information is also available on the planned and operated timetable so that capacity utilisation can be calculated. The nature of Britain's rail industry therefore means that there is a rich source of capacity utilisation and performance data which makes it an ideal subject for the questions posed by this thesis.

Furthermore, demand for rail capacity in Britain is increasing and as a consequence the rail network is becoming increasingly crowded. Official statistics show that by 2012 annual passenger kilometres were 57.3 billion (up 44% over the previous decade) and the amount of freight annually transported was 22.92 billion tonne kilometres (up 15.4% over the previous decade). However, this traffic growth was on a network that had shrunk in overall terms by 5.5% (from 16,652 kilometres in 2001/2 to 15,742 kilometres in 2011/12) (ORR, 2012a).

The increasing demand for travel can be accommodated within the existing rail network in a number of ways. The lengthening of existing passenger and freight trains is often seen as an attractive option. This is particularly true for passenger trains where the issue is over-crowding. Although, this solution can be relatively cheap, when only the provision of new rolling stock is required; costs can quickly increase if improvements to rail infrastructure are also required. In the case of passenger services this is often the lengthening of platforms. Infrastructure solutions can be both expensive and time consuming to implement. However, even ambitious schemes for existing lines may not be sufficient to cope with rising demand for capacity. The fact that the British Government, backed by many business leaders and pressure groups, is at the time of writing continuing with the proposed £43 billion[1] new High Speed Rail Line (HS2) line underlines the growing demand for increased rail capacity between London and other major centres of population in this country.

One key issue with this rise in capacity utilisation is the likely effect that this will have on timetable performance. As the network becomes busier, even small delays are magnified as following trains are themselves delayed. This so-called reactionary delay has a significant impact on the rail industry's attempts to deliver on-time services. Investigating the relationship between capacity utilisation and performance (specifically reactionary delay) is therefore a current and important theme. The growth of demand described previously increases the need to gain a better understanding of the impact that capacity utilisation has on performance. There is also a significant financial incentive to gain a greater understanding from both the point of view of the large sums of money that currently 'change hands' within the various performance regimes and also the risk associated with making the wrong decision over how to accommodate the predicted increases in demand.

An important theme of this thesis is the optimisation of capacity utilisation on the rail network. The difficulty and cost associated with the growing demand on the British rail network has already been explained. The rich source of data provided by the privatised rail system in this country has also already been discussed. An additional useful aspect of the British rail network is that it is also already subject to a congestion charge, called the Capacity Charge. Analysis of this charge provides a useful starting point to any analysis on charging for access to congested rail networks.

[1] Source : HS2Ltd, 2013a

## 1.2 Aims, Objectives and Methodology

### 1.2.1 Aims and Objectives of the Thesis

The aim of this thesis is to understand the impact that capacity utilisation has on the performance of a congested rail network. This relationship will be explored through the application of traditional regression techniques to data obtained from Network Rail for parts of Britain's East Coast Main Line (ECML). Previous work, which is discussed in the literature review contained in this thesis, suggests that there is an exponential relationship between capacity utilisation and reactionary delay. This is perhaps unsurprising given that as noted earlier reactionary delay is the secondary delay that services suffer due to a train in-front being delayed. Logically this 'knock-on' delay will increase at a greater than linearly rate as a network becomes busier.

The principle objective of the work is to:-

> *Understand the relationship between capacity utilisation and performance on a sample rail network and to use the results to make recommendations about the most appropriate charging mechanism for congested rail networks.*

This can be broken down into a number of distinct elements:-

1. The measurement of capacity utilisation on a sample rail network using a variety of methodologies.
2. The measurement of performance, and specifically reactionary delay, on a sample rail network.
3. The assessment of the relationship between capacity utilisation and reactionary delay for the sample network using established regression analyses techniques. A key objective is to determine which of the capacity utilisation measures considered provides the most 'effective' predictor of timetable performance.
4. An exploration of the role that 'other' factors play in the level of observed reactionary delay on the sample rail network.
5. The discussion of the transferability of the results to other congested rail networks.
6. An examination of possible charging mechanisms for congested rail networks, using the results from the sample network to illustrate the options discussed.

These objectives were developed following an extensive literature review of rail capacity utilisation and timetable performance and the charging for access to congested transport networks.

### 1.2.2 Methodology and Discussion of Aims

In order to investigate the relationship between capacity utilisation and timetable performance, data has been obtained from Network Rail for two parts of the southern portion of the East Coast Main Line (ECML) for the December 2009 to May 2010 timetable. As discussed later in the thesis, this part of Britain's rail network provides an ideal subject for the exploration of congested rail networks. The chosen timetable also provides a suitable data set for the analysis being carried out.

The capacity utilisation measures were determined following an extensive literature review, carried out to identify approaches used in previous relevant studies. In some cases these measures were adopted close to the original approach, subject to any necessary modifications due to the nature of the data set used for this thesis. In other cases, more major modifications were made to ensure that they were suitable for the analysis. For example, one important part of the analysis is the development and investigation of capacity utilisation measures which include junction moves rather than the standard approach which involves just link moves (i.e. the sections between nodes).

Performance data was obtained for the sample area. This was of two types. Firstly, reactionary delay for the relevant points on the sample network was collected and sorted. Secondly, lateness data was obtained. This provides information on how late traffic was when it entered the sample area and also how often a particular service operated. As will be seen, both types of data were used to examine the relationship between the capacity utilisation and timetable performance of the sample network.

The capacity utilisation and performance information were used to create a data set of the sample area of the ECML suitable for regression analyses. Standard econometric approaches based on previous relevant work and theoretical explanations were used to investigate the relationship. Standard 'success' measures were then used to explain which capacity utilisation measures were considered to be the most effective in explaining the relationship between it and timetable performance.

One important aspect of the work was to investigate whether other factors complemented or indeed provided a better explanation of timetable performance than capacity utilisation measures. Once again potential other factors were identified following a literature review. However, one important aim was to keep the relationship as parsimonious as possible in order to ensure that any theoretical tariff was both simple and transparent.

The potential transferability of any findings is clearly important. There is little to be gained from identifying relationships that only apply to a small sample area. However, the main aim is to establish relationships that will apply to other congested rail networks rather than necessarily all rail networks. In other words it is desirable that some of the detail of the findings is not lost by having to make the recommendations universally applicable.

The results of the analyses are then used to produce and compare possible congestion charge mechanisms. The merits of various approaches are discussed. One key element of the discussion is to consider whether alternative options to the current Capacity Charge approach are likely to be more logical. The Capacity Charge has applied since 2002 and levies a charge on all train movements on the rail network on the basis of the relationship between the volume of capacity usage and reactionary delay for given groups of train services. In part therefore this thesis provides an independent and alternative review of this charge.

## 1.3 Structure of the Thesis

Chapter Two provides the background to the importance of understanding the relationship between capacity utilisation and performance and the link to congested rail networks. A simple definition of rail congestion is provided and the substantial economic cost it causes is highlighted. The point is made that ‘hard’ solutions which involve the provision of new infrastructure are often time consuming and expensive to provide. The growth in demand for rail travel means therefore that ‘soft’ solutions which involve the optimal use of the existing infrastructure have become more and more important.

Chapter Three applies theoretical concepts to the issues discussed in Chapter Two through a literature review. The chapter begins with a discussion of the general features of traffic congestion. The principles of capacity utilisation measures are then explained with those that will be used in the analyses for this thesis identified. Finally, timetable performance is discussed and the findings of previous research into its relationship with rail capacity utilisation highlighted.

Chapter Four then applies these theoretical concepts to the actual performance of the timetable in Britain and provides a detailed account of how the relationship between it and capacity utilisation was used to develop the Capacity Charge in Britain.

Chapter Five outlines the methodology used to examine the relationship between capacity utilisation and performance (specifically reactionary delay). The steps in the regression analyses to explore the link between capacity utilisation and reactionary delay are explained. This includes the success measures which will be used to determine the most effective capacity utilisation and non-capacity utilisation measures. The methodology behind the creation of the data set is then discussed. Finally, the approach taken to consider the implications of the results is outlined.

Chapter Six explains how the data set was created to test the various capacity utilisation measures described in Chapter Three using the methodology outlined in Chapter Five. The reasons behind the choice of the sample network and timetable are explained.

Chapter Seven describes the results of the regression analyses and explains their significance. In particular the most effective capacity utilisation measure is identified. The reasons for the results are discussed using examples from the data set. Finally, the transferability of the results to other rail networks is considered.

Chapter Eight looks at the implications of the results presented in the previous chapter for the charging of congested rail networks. Potential tariffs are calculated and compared using the values obtained from the regression analyses. Alternative approaches are then considered with recommendations made on which are considered to be the most effective.

Finally, Chapter Nine provides some overall conclusions for the work; considers whether the original objectives described in this chapter have been met and makes recommendations about potential future work. Finally, the contribution of this thesis to a wider understanding of the issues covered are highlighted.

# Chapter 2
# Britain’s Capacity Challenge

## 2.1 Introduction

There is significant pressure on Britain’s rail infrastructure with more journeys now being made since 1927 (Thompson, G., Hawkins, O., Dar, A. and Taylor, M., 2012, p129). Passenger journeys have almost doubled since privatisation from 735 million in 1994-95 to 1.6 billion in 2011-12 and rail freight has expanded by over 60% to 21.1 billion tonne kilometres per annum (Transport Select Committee, 2013, p5). This pressure is expected to worsen and “some of the country’s key rail routes are forecast to be completely full in peak hours in the next 20 years” whilst the volume of rail freight on the network is expected to double by the year 2030 (Department for Transport, 2012b, pp10-11).

This growth undoubtedly puts additional strain on Britain’s rail network. Infrastructure cannot easily or cheaply be expanded. At the same time, Network Rail is under pressure to improve the performance of the network and reduce costs. It is clear that the industry is faced with a number of difficult choices. For example, relatively recently Network Rail reported that the West Coast Main Line (WCML) despite its modernisation a few years earlier was a comparatively ‘poor’ performer and the introduction of further services to cope with rising demand was likely to put even more pressure on reliability (Department for Transport, 2012b, p13). This problem is referred to by Khadem-Sameni, M., Preston, J. and Armstrong, J. (2010) as Britain’s Capacity Challenge.

This chapter provides the background to why it is important to understand the relationship between capacity utilisation and performance and highlights the link to rail congestion. Firstly, a simple definition of rail congestion is provided. Secondly, its substantial economic impact is briefly described. Finally, the difficulty of finding appropriate solutions to the Capacity Challenge are outlined in some detail. These difficulties are used to support one of the conclusions of the Eddington Report that there should be “a focus on the performance of the existing network, particularly where capacity is stretched, as demonstrated, for instance, through congestion or unreliability” (Eddington, 2006, p3).

## 2.2 A Simple Definition of Rail Congestion

Rather confusingly although there is an official definition of congested infrastructure this will not be adopted for this thesis. Network Rail is legally obliged[2] to declare parts of the network 'congested ' when certain conditions are met. One of the conditions is that Network Rail, after coordination with all parties requesting access to capacity, has not been able to satisfy all requests adequately. Once a part of the network is declared congested, Network Rail must give notice of this (through their annual Network Statement), undertake a capacity assessment and develop a capacity enhancement plan where one is not already in place.

However, this official condition of congested infrastructure is clearly in theoretical terms a state of 'scarcity' rather than 'congestion'. This is clarified by the following definitions. 'Scarcity' occurs during the timetable development process when due to capacity limitations "use of a particular slot by one train operator leads to the inability of others to obtain their desired slots" (Johnson and Nash, 2008, p53). In contrast, 'Congestion' for the purposes of this thesis refers to a state in the relationship between capacity utilisation and performance[3] during the actual operation of a timetable. Specifically it refers to a point when the level of capacity utilisation begins to have a detrimental impact on timetable performance.

Although, the emphasis of this thesis is on rail 'congestion' the two concepts clearly have a common basis and 'scarcity' will be referred to elsewhere as appropriate. Many of the conclusions relating to congestion also equally apply to scarcity.

Interestingly, the Congested Infrastructure Declaration has been rarely applied by Network Rail. The 2014 Network Statement notes that two declarations of Congested Infrastructure have been made since 2008 and no other areas of rail network were being declared congested by Network Rail

[2] This requirement is contained in The Railways Infrastructure (Access and Management) Regulations of 2005 which were amended in 2009. These are two statutory instruments which implement a number of EU directives under UK law.

[3] A detailed explanation of this relationship will be provided in Chapter Three.

Network Rail, 2012e, pp47-48). Both were for relatively small sections of the network[4].

This lack of declarations has drawn some criticism from the rail industry. Alliance Rail Holding Ltd (a potential open access operator with aspirations for paths on both the WCML and ECML) wrote to the ORR in February 2013 with their response to Network Rail's Strategic Business Plan for Control Period 5 (2014 to 2019). In this they specifically referred to the issue of congested infrastructure saying that:

> "Network Rail is seeking a number of very large scale enhancements that do not address train path capacity. For example despite the significant sums invested and due to be invested on the WCML, Network Rail will not sell known validated paths. At the same time Network Rail refuses to declare parts of the Network formally congested"

asking elsewhere in the letter "will the infrastructure capacity enhancements (proposed for the rail network) actually deliver capacity or will the outputs be used for performance robustness?" (Alliance Rail Holdings Ltd, 2013, pp1-2). This final point once again underlines the fact that the rail industry is faced with a number of difficult choices.

## 2.3 The Cost of Rail Congestion

In simple terms an efficient rail network is important to the success of the British economy. As concluded by the Eddington Transport Study (2006, p3) "transport matters for the economic performance of countries and regions".

The cost of congested rail infrastructure to the British economy is believed to be substantial. There is a link between this cost and capacity utilisation and rail performance. Edward Leigh, MP (Parliament, 2008) and at the time Chairman of the Public Accounts Committee, summed up these elements in his response to a report on how delays to rail passengers could be managed more effectively. He said:

> "Rail passengers pay handsomely to travel on trains (£5.1 billion in fares in 2006-07) and yet, through incidents on the network, are still suffering expensive delays (£1 billion in lost time in 2006-07).

[4] One was for a strategically important freight route in Scotland principally used by coal traffic. The other was the route between Reading and Gatwick Airport.

> Performance has returned to the levels that existed before the 2000 Hatfield derailment, but increasing congestion on the network means that the consequences of an incident in terms of disruption are magnified".

In other words greater capacity utilisation means that any initial delay will be amplified and performance will suffer leading to increased costs. The increasing demand for rail travel and the associated costs of delays means that the importance of understanding rail infrastructure congestion therefore cannot be understated.

## 2.4 Approaches to the Capacity Challenge

### 2.4.1 Overview

One of the key conclusions of the Eddington Transport Study (2006, p3) was that there was no single solution to transport problems since transport needs vary so widely. Khadem-Sameni, M., Preston, J., and Armstrong, J. (2010, p5), in their conference paper on Britain's Rail Capacity Challenge, divided solutions into 'hard' and 'soft' approaches. 'Hard' solutions involve enhancing existing railway infrastructure or providing additional infrastructure in for example the provision of new railway lines. 'Soft' solutions involve making better use of existing capacity through timetable optimisation and demand management. The latter involves 'pricing' strategies, including the concept of congestion charging which forms a key element of this thesis. In order to understand why 'soft' strategies are of growing importance as a solution to finite capacity it is necessary to examine both sets of approaches.

### 2.4.2 'Hard' Solutions

Infrastructure or 'Hard' solutions can be divided into four basic categories. These are:-

- Increase in train length / width / height.
- Localised Infrastructure Enhancements
- Line Modernisation
- New Railway Lines.

Making trains longer, wider or higher means that more passengers or freight can be carried without increasing the number of services on already crowded networks. However, substantial increases to a trains' carrying capacity are difficult to achieve without expenditure on changes to the existing infrastructure. For example, for passenger trains platform lengths often have to be increased as well as new carriages purchased.

There is also only so much that this approach can achieve.  The strategic case for the proposed new HS2 line (HS2Ltd, 2013b, p12) notes that: “additional seats are being provided by lengthening trains and for a while this will address the problem of growing demand. But this will not address the problem beyond the next 10-15 years”.

Localised infrastructure enhancements are intended to improve the capability of the existing network. However, as noted by the Department for Transport the scale of expected future demand on key routes means that relatively easy incremental changes such as “minor local layout and signalling modifications” will not be sufficient by themselves and that incremental changes such as grade-separation at junctions which can provide substantially more capacity are “progressively more costly” and that “land availability makes additional surface running lines in urban areas prohibitively expensive in most cases” (Department for Transport, 2007b, p12).

One example of a large scheme is the Reading Station Area Redevelopment scheme. Network Rail’s Enhancement Plan says that it is “designed to deliver significant capacity and performance improvements throughout the area for GWML (Great Western Main Line) and cross country passenger and freight services” (Network Rail, 2011b, p17). However, it has a listed expenditure of £161 million. Additionally completion of the first element of the scheme was December 2010 but the final date for the scheme was not due until April 2015.

A major challenge is the overall age of Britain’s rail network and the need to modernise it. The rail network at the time Network Rail was formed was not only old but had suffered from almost 50 years of underinvestment (Network Rail, 2013a, p13). In some cases targeted enhancements are deemed insufficient to deal with capacity issues and modernisation of an entire route is considered the only sensible solution. The upgrade of the West Coast Main Line completed in 2008 provides an excellent example of this. The modernisation of the core route between London and Glasgow and its key divergences to Birmingham, Manchester and Liverpool has been the largest rail project to date in Britain. The improvements dealt with significant capacity constraints, permitting more frequent services and the reduction of journey times.

However, these improvements came at a price. The final cost of the project was estimated to be £9.9 billion compared with an initial estimate of £1.5 billion (Butcher, 2010, p16). The upgrade took eight years to complete (following the 37 months required to achieve planning approval, Business Infrastructure Commission, 2013, p14) and resulted in significant disruption to journeys due to the need for major 'blocks' of the existing railway to carry out the engineering work. Furthermore, in July 2013 it was reported that Network Rail had rejected Virgin Trains' bids for new services from London to Blackpool and Shrewsbury on the basis that the WCML route "could not cope with more traffic" (BBC News, 2013). This rejection of an access request on capacity and performance grounds, just five years after the completion of the WCML modernisation project; underlines the difficulty of accommodating the growing demand through infrastructure solutions alone.

The ultimate infrastructure investment is the construction of new railway lines. The current proposal to construct a new high-speed line between London, the Midlands and the North (HS2) demonstrates the difficulties, very long timescales and huge expense associated with such an undertaking. At the time of writing there is an on-going debate about whether the benefits that will be obtained from the new line justify the huge costs. The Government and the line's supporters are certainly of the opinion that only by providing the additional capacity that the new line will create, can the predicted growth in rail travel be accommodated at an acceptable cost and with reliable performance.

The new line will be 351 miles long and is the first new railway north of London for 120 years. The first phase from London to Birmingham is currently planned to open in 2026. The second phase which extends the line to both Manchester and Leeds is due to open in 2033. (HS2Ltd, 2013a).The decision to proceed with the new line was taken in 2012 meaning a 21 year timescale until the full benefits of the scheme are achieved. The benefits of the new line listed by its promoters include: 'freeing up' space on the existing rail network (to accommodate for example the growth in freight traffic); faster and 'better' journeys between cities; economic growth including employment creation and reduced emissions (HS2Ltd, 2013b)

However, there is a very significant cost associated with building new railway lines. In addition to construction costs, substantial land purchase and compensation schemes will be required representing a significant proportion of the overall cost. The current total budget for the HS2 line is £42.6 billion (including £14.4 billion of contingency). (HS2Ltd, 2013a).

### 2.4.3 New Technology

Before moving onto a discussion of ‘soft’ solutions it is worth noting the part that new technology plays in addressing Britain’s Capacity Challenge.

Investment in new technology can reduce the level of congestion on the network by maximising the effectiveness of ‘hard’ solutions. For example, investment in modern signalling systems can permit trains to travel safely and efficiently closer together than presently. By increasing the capacity of the railway line in this way, the level of congestion will be reduced.

The ‘next generation’ of signalling in-fact involves the introduction of ‘in-cab’ signals. In Europe this is being developed as the European Train Control System (ETCS) which forms part of the European Rail Traffic Management System (ERTMS). Computerised signalling systems in the trains themselves can increase the available capacity since the distance between trains will be continuously evaluated and the particular braking and accelerating characteristics of each train will be constantly monitored. The introduction of ETCS reduces the permissible safe distance between trains whilst allowing higher speeds. The ERTMS programme estimates that increases in available capacity will be as much as 40% (ERTMS Website, 2013, p1). Although, the overall cost will be very considerable there is a belief that the introduction of ERTMS is essential. “It will mean that capacity usage of our crowded rail network can be optimised” (Department for Transport, 2011).

Investment is also being made in new types of rolling stock. The Intercity Express Programme (IEP) provides an example of a very substantial investment in this. IEP is intended to replace the ageing intercity trains, particularly on the GWML and the ECML. The trains will be lighter and more reliable than the existing rolling stock, meaning that less track maintenance will be required and fewer train performance issues are likely. This suggests fewer associated primary delay incidents will occur. The trains will be faster, have better acceleration and more seating capacity than the rolling stock they are intended to replace. Sir Andrew Foster in his 2010 Independent Review noted that one of the high-level critical success factors of the IEP programme was that the new trains make “best use of available route capacity” (Foster, 2010, p9).

However, newer and better rolling stock is an expensive and long term investment.  The Department for Transport and Hitachi (the trains’ manufacturers) defended the £5.2 billion investment in an article on the Guardian newspaper’s website in December 2013 saying that “The government's Intercity express programme is a multi-billion pound project that must be delivered if we are serious about rolling out a rail network fit for the 21st century” (Hammond, S. and Dormer, A. 2013).

Investment is also being made in the fields of maintenance and renewal for developing new techniques which minimise for example the amount of time needed for disruptive possessions. However, the benefits of new technology will only be achieved with the investment of sufficient time and money. Network Rail have themselves admitted to under-investment in research and development in previous years but are now “rapidly making up for lost time” and by 2019 “will be investing more per year than other comparable British companies” (Network Rail, 2013a, p20).

### 2.4.4 ‘Soft’ Solutions

The previous sections highlight the difficulty of addressing the growing demand for capacity through infrastructure investment alone. Abril, M., Barber, F., Ingolotti, L., Salido, M.A., Tormos, P. and Lova. A. (2008, p774) note that “capital expansion is a very costly means of increasing capacity. A more effective solution is to manage the existing capacity more effectively”. This increased emphasis on better management of the existing infrastructure is echoed by the McNulty Report (2011, p11) which said there “should be an end to ‘predict and provide’ in the rail sector and there should be a move towards ‘predict, manage and provide’ with a much greater focus on making better use of existing capacity”.

These so called ‘soft’ solutions can be divided into three basic categories. These are:-

- ‘Better’ Timetables
- ‘Better’ Engineering Access
- Demand Management .

Timetables that use capacity more efficiently is an important part of this thesis and will be covered at greater length in subsequent chapters. ‘Better’ Engineering access refers to the concept, referred to previously, of less disruptive possession being taken of the network for the necessary maintenance, renewal and enhancement work. More innovative possession solutions in this field means more available capacity for traffic thereby helping to reduce the level of congestion.

It is worth noting here though that one of the roles of the ORR is to oversee the “fair and efficient allocation of capacity” (ORR 2004b, p17). For example, this is reflected in their published strategy for CP5, where they state their goal for 2009-14 is that “the main-line industry has in place arrangements to achieve the best use of capacity on the network” (ORR 2009, p24). Network Rail themselves has an objective contained in Part D of the Network Code[5] of sharing capacity “in the most efficient and economical manner" when making timetable decisions (Network Rail, 2014b, p31).

Demand Management itself covers three basic ideas. These are:-

- Pricing the end customer.
- The Scarcity Pricing of Paths.
- The Congestion Pricing of Paths.

The idea of customer demand management is to reduce pressure on capacity in the peak period. The Department for Transport considers that “systems and incentives need to be put in place to make better use of assets, so that we encourage existing customers to modify their usage of the railway towards quiet, off-peak periods when there are empty seats, empty wagons and even spare train paths available” (Department for Transport, 2007b, p20). Encouraging passengers to use services at less congested times where possible through the use of differential ticket pricing is a wide spread approach.

Differential ticket prices already apply to the British rail network with ‘Advance’, ‘Off-Peak’ and ‘Anytime’ tickets currently available depending on the nature of the journey[6]. However, Whelan and Johnson (2004) found that the differential between peak and off-peak fares needed to be substantial to affect over-crowding with a combined strategy of increased peak fares and reduced off-peak fares. There is of course a delicate balance required with any pricing strategy. Increasing peak fares by too much risks encouraging customers to switch transport modes and leads to potentially greater road congestion. This is contrary to current government policy and the level of peak fares that franchised operators can charge are in-fact regulated. Reducing off-peak fares by too much could however reduce the income of

[5] “The Network Code is a common set of rules and industry procedures that apply to all parties who have a contractual right of access to the track owned and operated by Network Rail” (ORR , 2014).

[6] Source : National Rail Enquiries, 2013.

train operators to unacceptable levels. Furthermore, Network Rail whilst noting that there might be some scope to spread passenger demand through ticket pricing acknowledged that “these opportunities are likely to have already been exploited by TOCs” (Network Rail, 2009a, p26).

A seemingly logical step is to levy a charge for access to scarce paths to ensure that they are allocated in the most efficient way possible. Specific scarcity charges currently do not apply to the UK rail network and concern has been expressed in the past that their introduction might allow the Infrastructure Owner to levy monopoly rents where the network is congested (Gibson, S., Cooper, G., and Ball, B., 2002). Nash, C., Johnson, D. and Tyler, J.  (2006) report however that, at the time of their research, a number of other European countries (e.g. Germany) had applied ‘scarcity’ surcharges to ‘busy’ sections of track. There has also been a great deal of academic interest in the form that a charge should take.

Three basic forms of scarcity charges have been identified:-

- An auctioning process
- Charging the Short Run costs
- Charging Long Run costs.

The concept of auctioning scarce timetable slots appears on the face of it an attractive proposition. Slot allocation is determined on the basis of willingness to pay and the infrastructure owner is able to theoretically maximise revenue which can in turn be invested in enhancing the network. However, there are substantial difficulties to overcome. These include the determination of which slots to be auctioned; the need for a complex iterative process to ensure that the paths obtained are compatible and the need to ensure that ‘paths’ required on a social welfare basis are not lost due to a desire to maximise revenue. These issues are recognised by Nilsson (2002), amongst others, who did not believe that they were insurmountable.

However, Gibson (2003) and Thomas and McMahon (2005) both make the point that whilst the auctioning of scarce capacity is a market-based approach; in the UK the allocation of capacity broadly follows an administrative approach where it is determined by a third party (i.e. the ORR in their role of approving access rights).  Nash, C. Johnson, D. and Tyler, J. (2006) do suggest though that an auction process could be used to allocate spare marginal slots in the UK after for example the passenger franchisees had fulfilled their obligations, provided considerable care was taken with the allocation of compatible slots.

Nash, C. Coulthard, S. and Matthews, B. . (2004) explored the principle of charging for scarcity with a case study of the Transpennine route. They concluded that the appropriate charge is the social opportunity cost of the train ‘forced off’ the network by another train due to the lack of sufficient capacity.

Johnson and Nash (2008) modelled the value of existing peak and off-peak franchised services for each direction on the East Coast Main Line and the cost of replacing them with open access paths using the PRAISE software. Their results seemed to confirm the view that existing variable charges for key routes where capacity is scarce were set at much too low a level. They concluded that the imposition of scarcity charges based on the value of slots to the franchisee was both feasible and socially beneficial. However, they recognised that further work was required to determine what form the scarcity charge should take. One issue with the use of short-run incremental costs alone is that they do not meet with one charging objective of the ORR which is to ensure that the structure of charges provide incentives for not only efficient utilisation but also development of the rail network (Thomas and McMahon, 2005).

The third approach therefore is to identify those sections of infrastructure where capacity is scarce and charge the long run incremental cost of expanding capacity. The attraction with this approach is that supply is made to match demand. There are however a number of issues. Firstly, as previously described there are very long lead times associated with infrastructure enhancement works. There is therefore a need to accurately anticipate demand some way into the future if this approach is to be effective. Secondly, the cost of expanding capacity can vary enormously depending on the exact proposal being considered. Furthermore, as Turvey (2000) points out, the creation of additional capacity on a route may produce large ‘blocks’ of additional capacity over time thus leading to the problem of ‘lumpy’ investment which in turn can create confusing price signals. Gibson (2003) notes that the value of any additional paths created is often unlikely to match the significant cost of any infrastructure enhancement to relieve capacity bottlenecks. Thirdly, there is the issue of how to levy a charge when a mix of operators both current and potential stand-to-gain from any increase in available track capacity.

In contrast, a congestion charge called the Capacity Charge is levied on the British rail network. It has been in place since 2002. Gibson, S, Cooper, G., and Ball, B. (2002, p342) referred to the Charge as the “first time that an infrastructure manager has sought to introduce such a highly disaggregated congestion-related charge across a rail network”. Nash (2005) notes that Infrastructure Managers in other European Countries (e.g. Germany) in contrast adopted a much simpler approach based on applying surcharges for use of specific congested links or nodes. Due to the relevance of congestion charging to this thesis, the history, development and use of the Capacity Charge itself will be covered in detail in Chapter Four.

## 2.5 The Role of Different Inputs to the Process

There are therefore a number of widely different approaches to tackling Britain’s Capacity Challenge. However, the choice of which to adopt is heavily influenced by a number of inputs to the process.

Firstly, through legislation and its role as a major funder, Government policy has a huge influence on how rail congestion is addressed in Britain. The 2004 White Paper ‘The Future of Rail’ included the statement[7] “the Secretary of State for Transport will take responsibility for setting the national-level strategic outputs for the railway industry, in terms of capacity and performance” (quoted in Department for Transport, 2008, p3). In 2007 the Government’s strategic policy towards rail transport in Britain was clarified in its White Paper “Delivering a Sustainable Railway” (Department for Transport, 2007a). It states that “safety, reliability and cost are permanent priorities for the railway. But increasing capacity is the most urgent investment need – to accommodate record passenger numbers, allow rail to contribute to low-carbon economic growth, and move towards the service quality that more exacting consumers increasingly demand”. (Department for Transport, 2007a, p13).

Detailed policies are developed using the Network Modelling Framework (NMF) which is a detailed strategic forecasting and appraisal model. This was developed using a co-operative approach with the industry’s stakeholders. The NMF’s purpose is to support decision making by the government and the ORR. Its inputs include demand, timetable assumptions and an assumed level of fares. These are used to calculate metrics which include capacity utilisation, performance, crowding and operating costs.

The output from the NMF feeds into the High Level Output Specification (HLOS) and influences the Statement of Funds Available (SoFA), each of

[7] White Paper (the Future of Rail) 15 July 2004 section 3.2.6

which relate to a specific five year control period. The requirement for the government to produce these two documents was contained in the 2005 Railways Act. The object of the HLOS is to inform the ORR, and the rest of the rail industry, about the level of capability (including the capacity and performance) of the railway that the Government wants to see. The object of the SoFA is to detail the amount of public funding that the Government intends to make available to enable the industry to deliver the outputs set out in the HLOS.

The HLOS and SoFA for 2014 to 2019 reveals that the Government expects passenger demand to grow by 16% and freight by 23% during this period (Department for Transport, 2012a, p2) and as a consequence “the Secretary of State wants to see a significant increase in the carrying capacity of both the freight and the franchised passenger railway”(Department for Transport, 2012a, p6). Although £5.2 billion was committed to enhancing the infrastructure (Department for Transport, 2012a, p2), the Government wished to see the cost of operating the railway reduce by £3.5 billion by 2019 (Department for Transport, 2012a, p5) and also an improvement in the performance of the railway (Department for Transport, 2012a).

As part of the delivery process Network Rail are obliged under their licence conditions to produce the Route Utilisation Strategies (RUSs). These are produced through extensive consultation with industry stakeholders. Network Rail states that they “seek to balance capacity, passenger & freight demand, operational performance and cost, to address the requirements of funders and stakeholders” and that “Network Rail will take account of the recommendations from RUSs when carrying out its activities. In particular they will be used to help inform the allocation of capacity on the network” (Network Rail, 2013e).

The RUSs first examine what the system can do now (supply) and what is expected of it (demand) and any gaps between the two are then identified. Recommended options are then presented as a ‘menu’ “from which funders may select the future outputs of the network” (Network Rail, 2009a, p31). The Route Utilisation Strategies because they are intended to give a comprehensive review of how to balance likely supply and demand on a route, therefore provide a valuable resource on capacity issues for Britain’s rail network.

## 2.6 Summary

This chapter has discussed Britain’s Capacity Challenge. A growth in demand for access to the rail network has to be balanced with a need to maintain reliability and minimise costs. One important approach in seeking to meet this challenge is the effective use of existing capacity. In order to achieve this it is necessary to understand the relationship between capacity utilisation and performance and the nature of rail congestion. The next chapter therefore considers these three aspects through the use of a literature review.

# Chapter Three
# Capacity Utilisation, Performance and Rail Congestion

## 3.1 Introduction

The concepts of the capacity utilisation and performance of rail networks have both attracted a great deal of academic interest. This reflects their importance. Network Rail in their 2013 publication ‘A Better railway for a Better Britain’ lists them as two of the three key challenges they face, with the other being cost (Network Rail, 2013a, p5). This chapter explains these terms through a literature review and how they can lead to rail congestion. Methods of measurement are discussed and an explanation given on how they will be used in this thesis. The chapter however begins with a discussion on the general features of traffic congestion.

## 3.2 Traffic Congestion

Button (2004) refers to congestion as a consequence of transport infrastructure in the short run having a finite capacity. Goodwin (2004, p7) notes that the general feature of congestion is that users affect each other’s freedom of movement, defining it as “the impedance vehicles impose on each other....... in conditions where the use of a transport system approaches its capacity”.

An important point is there is more than one type of transport congestion. Vickrey (1961, p251), in a very influential paper, listed six different types of congestion and noted that these were often encountered in various combinations

These are:-

- Single Interaction.
- Multiple Interaction
- ‘Bottle-Neck’
- ‘Trigger-neck’
- Network and Control
- General Density.

‘Multiple interaction’ is associated with high volumes of traffic and Vickrey refers to the speed-flow relationship when discussing this type of congestion.

However, he recognises that congestion can also occur in light traffic conditions. ‘Single interaction’ means only two vehicles are involved but they are travelling too close together resulting in the following vehicle being forced to brake. Vickrey suggested that overall delay would be much higher for multiple interaction congestion than single interaction congestion. However, it is important to note that the relationship between capacity utilisation and performance is therefore defined by the size of the ‘gap’ (or ‘buffer’) between successive vehicles rather than the actual volume of traffic. This concept will be returned to later in the chapter.

‘Bottle-neck’ congestion is where one part of the route has less capacity than that available in previous and subsequent sections. As long as the flow does not exceed the capacity through the bottle-neck there will be little delay. However, if traffic continuously exceeds the capacity through the ‘bottle-neck’, queues will begin to form leading to substantial delays. The concept that the capacity of a rail ‘bottle-neck’ (or critical section) defines the potential capacity of the surrounding network will also be returned to later in this chapter. ‘Bottle-neck’ congestion can lead to the associated ‘trigger-neck’ congestion which is where the queues begin to interfere with traffic not intending to use the actual ‘bottle-neck’ itself. The observation that reactionary delays can propagate quite widely, particularly in highly connected and high density timetables will also be returned to.

‘Network and Control’ congestion describes levels of flow at such a level that interventions are necessary to regulate the flow of traffic and avoid ‘grid-lock’. In the case of road infrastructure these measures include stop signs, traffic lights and routing limitations. In a rail traffic context this includes Network Rail’s signalling staff making decisions about the priority given to different services. The type of interventions that can be implemented and how successful they are will obviously reflect the size, type and duration of the original primary incident; the nature of the infrastructure both on the affected route and any potential diversionary routes and the nature of the train services affected.

Finally, general density congestion describes the situation where traffic is at such a high level across the network as a whole that delays will occur at multiple points. This again suggests a strong relationship between capacity utilisation and performance.

All of Vickrey’s definitions of congestion can therefore be applied to a rail context.

## 3.3 Capacity Utilisation

### 3.3.1 General Principles of Capacity Utilisation

Examination of the literature makes it clear that rail capacity utilisation has a number of basic principles.

Firstly, “capacity as such does not exist. Railway infrastructure capacity depends on the way it is utilised” (UIC, 2004, p1). It will be seen that the type and frequency of rail traffic has a huge influence on how much ‘spare’ capacity a rail network has. This means the capacity of rail infrastructure cannot be determined without first making some decisions about how it is utilised.

Secondly, Krueger (1999, p1195) observes that capacity utilisation can be expressed in a variety of ways including the tonnage moved, the number of trains per day and available track maintenance time. This causes problems, since as noted by Krueger, many definitions are incompatible with each other. The reason that there are different definitions of capacity utilisation is that the metric chosen will depend on the issue being considered. In this thesis, for reasons that will become clear, rail capacity utilisation is talked of in terms of timetabled train paths.

Finally, “railway capacity .... is an elusive concept that is not easily defined or quantified” (Burdett and Kozan, 2006, p617). The reason for this is the numerous inter-acting factors that influence the capability of a rail network particularly where it is complex. It will be seen that a variety of different approaches have previously been proposed for measuring rail capacity utilisation which vary from the simplistic to the very complex. The choice of which approach to take of course depends on the objectives of the analysis. In this thesis, again for reasons that will become clear, rail capacity utilisation will be considered in a fairly high-level way.

### 3.3.2 The Capacity Balance

UIC (2004) explains the inter-action of four key factors in their well-known diagram ‘The Capacity Balance’. This is reproduced as Figure 3.1.

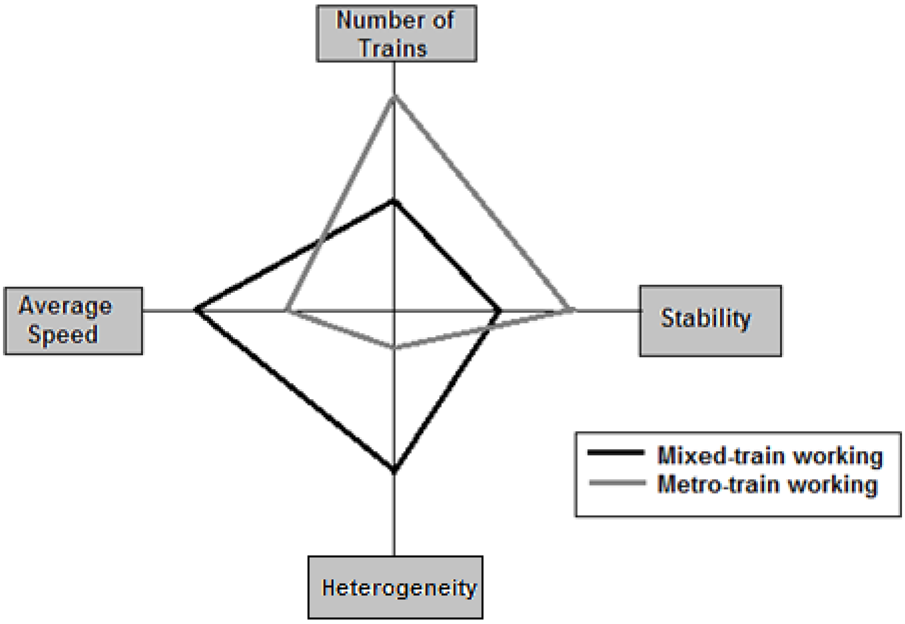


**Figure 3.1** – The Capacity Balance (UIC, 2004, p3).

It can be see that the 'Capacity Balance' is governed by four key factors: the number of trains, the stability, the heterogeneity and the average speed. Two types of train working are shown. Each has a 'chord', the length of which illustrates the overall available capacity. Capacity utilisation is defined as the position of the chord on each of the four axes. It can be seen that the chords for the different types of train working have different positions on each of the four axes.

Metro-train working, characterised by frequent services stopping at the same stations which are located relatively close together; has a high number of trains and stability. The low heterogeneity and average speed means that the metro-type timetable is fairly stable (or resilient) to performance issues. A delayed service can just take the path of the following train without delays becoming magnified and transmitted over a wider network.

It can be seen that the capacity utilisation of mixed –train working is very different from that of metro-train working. A number of trains are sacrificed due to the mixed nature of the traffic. For example, fast trains will begin to 'catch' slower trains reducing the size of the gaps in the timetable to operate other services. The number of trains and the heterogeneity of the timetable are key themes in this thesis. It will also be seen that the stability of the timetable is more likely to be affected by a heterogeneous timetable than a homogeneous one. Finally, the average speed of mixed-train working is likely to be higher than that of metro-working; however that in itself will reduce capacity due to the greater level of acceleration and deceleration.

### 3.3.3 Infrastructure Factors

It is also important to note that the "the basic parameters underpinning capacity are the infrastructure characteristics themselves" (UIC 2004, p2). For example, an intra-urban metro line will have very different infrastructure to an inter-city mixed traffic line.

A number of researchers have listed various elements that contribute to the potential capability of a rail network. These include Krueger (1999) and Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008). The contents of these lists do however depend on the nature of the rail infrastructure that the author is interested in. For example, Krueger (1999) describes his work on a capacity model he developed for the railways in Canada. These are largely single-track railways with intermediate passing points which cater for predominantly freight long-distance rail traffic. His list of infrastructure factors (p1196) reflects this:-

- Length of the subdivision (roughly 125 miles per sub-division).
- Average spacing of passing points.
- How equally spaced passing points are.
- Percentage of double-track line.

Secondly, the contents depends on how they are intended to be used. Krueger's list was for a specific model he had developed. Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008), as part of a review of different approaches to measuring capacity utilisation, provided a more general list. They suggest (pp777-778) that infrastructure parameters include:-

- The presence of 'single' or 'double' tracks.
- The signalling system.
- The  nature of the infrastructure e.g. gradients.
- Speed Limits.

### 3.3.4 General Approaches to Measurement

Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008, pp780-781) note that the measurement of capacity utilisation can be divided into three basic approaches. These are analytical methods, optimisation methods and simulation methods. They vary in complexity, realism and how general or specific they are.

Analytical methods are designed to establish capacity utilisation through theoretical formulae or algebraic expressions. They can vary between simple formulae with very few variables to significantly more complex models. The former are more likely to produce general models of capacity utilisation whilst the latter with a much greater degree of complexity may be very specific to particular locations or scenarios. One example of a complex mathematical model is the one produced by Krueger (1999) for Canadian Railways.

Optimisation methods are designed to address capacity utilisation issues and are based on the use of various techniques that examine the impact of adding additional traffic to already 'saturated' timetables. There has been a great deal of research into this approach. For example, Oliveira and Smith (2000) model the timetable as a special case of a job-shop scheduling problem with trains being treated as resources. They use this approach to develop a hybrid algorithm. Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008, p781) note that optimisation methods generally provide much better solutions to capacity problems than the simpler mathematical approaches.

Simulation methods are the most realistic but as Khadem-Sameni, M., Preston, J. and Armstrong, J. (2010, p3) note they are data intensive and computationally difficult. Sophisticated off-the-shelf software (e.g. Railsys) is used to produce a very detailed analysis of the operation of rail infrastructure. In his review of timetable planning for his 2008 PhD thesis, Watson (2008) suggested that at the time too little advantage was taken of these new approaches in Britain. Since then interest has grown in the use of Railsys in this country and it has now become a standard part of Network Rail's capacity planning 'tool-kit' (Network Rail, 2013d). There have however been some issues with its early use though. For example, MVA Consultants (2010) in a lessons learnt exercise  identified problems with the interpretation of the output from a Railsys study of a new West Coast Main Line timetable by non-technical 'customers'.

A simple theoretical formulae approach is the one adopted for this thesis. This is because one objective of this thesis is to establish whether the simple methodology used to calculate the current Capacity Charge can be improved upon. The use of a simple approach also maximises the likely transferability of any findings.

### 3.3.5 Sectional Running Times

A simple potential way to calculate capacity utilisation is to examine the transit time between two points. In Britain, the transit times between two important locations are referred to as Sectional Running Times (SRTs). These are calculated by Network Rail using a variety of approaches including the actual timing of trains and computer simulation and are then agreed with Train Operators. Network Rail then rounds the SRTs to the nearest half-minute. They are potentially an important input to capacity calculations as they reflect the infrastructure parameters of the section in question (e.g. speed limits, track curvature and gradients) and the operating characteristics of the traffic using it (e.g. acceleration and deceleration times and top speeds).

However in practice, a line's capacity will generally not be determined by the SRT. This is because the overwhelming number of sections have intermediate signals and it is these that play the major role in determining the capacity utilisation of a railway line

### 3.3.6 Headways

An important step is therefore to consider the role of signalling in determining the level of potential capacity utilisation. As noted earlier in the Chapter, one of the infrastructure factors listed by Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008) is the applicable signalling system.

The role of signals is to keep trains a safe distance apart. They work on the principle that only one train can be in a track section or 'block' at any one time. Clearly, placing signals closer together will increase capacity as the transit time of each block section is reduced. However, there is a limit on how close signals can be placed together. Aside from the cost consideration there is the issue of a driver responding in time to a red signal. Multi-aspect signals therefore use yellow lights in the sequence to alert train drivers that they are approaching a red light section. This approach allows drivers to regulate their speeds in a more efficient way increasing the number of trains that can be safely accommodated on a network. As described in the previous chapter the next generation of signalling removes 'fixed' signals altogether and introduces 'in-cab' signals which further increases the capacity of a line.

A key component in the calculation of capacity utilisation is therefore the permissible minimum gap or 'headway' between successive trains. At this stage it is necessary to divide them into technical and planning headways. Technical headways are the actual calculated minimum gaps that apply to a specific ‘block’ section of track. In order to calculate capacity utilisation as accurately as possible, these are the headways that would be used. However, the values will naturally vary between adjacent sections and for timetable planning purposes a common value is usually applied to groups of similar sections. Furthermore, whilst technical headways will be calculated in seconds; planning headways are commonly calculated to the nearest half-minute.

The difference between the technical and planning headways can be surprisingly large on a route. For example, the RUS for the ECML (Network Rail, 2008a, p197) showed in a chart of the Down[8] evening peak capacity utilisation, 80% planning headway utilisation for Welwyn Viaduct compared with a 35% technical headway utilisation.  The differences produced by the two types of headway  has led to some robust comments about which Network Rail should use to calculate capacity utilisation (e.g. Alliance Rail Holding Ltd, 2012).

There is also the issue that headways tend to be calculated on a ‘green-to-green’ basis which is the minimum gap between trains that would mean the following train always receives a green aspect. This has led to calculated capacity utilisation figures which exceed 100% for example at peak hours on the approaches to some of the London stations (Arup, 2013). In other words, for all the traffic to be accommodated it is necessary to plan them so that the trains are expected to receive yellow aspects.

Whilst technical headways can be difficult to obtain, planning headways are published by Network Rail in their annual Timetable Planning Rules. Table 3.1 shows an extract for the Up Direction for the southern portion of the East Coast Main Line. These are shown together with sample speed limits for the relevant sections obtained from the Sectional Appendix for the route, another document produced by Network Rail.

The portion of the route shown consists of a large number of signalling sections. However, it can be seen that these have been consolidated into a standard headway (4 minutes) with a number of exceptions giving a range between 3 minutes and 5 minutes. Clearly, those sections where trains can

[8] By convention the direction towards London is referred to as the ‘Up’ and away from London the ‘Down’.

be planned 3 minutes apart will have a much greater capacity than where the headway is 5 minutes. The slow lines generally have a greater headway than the fast lines (where there are both).

**Table 3.1** Planning Headways and Sample Speed Limits for the Southern Portion of the ECML (Sources: Network Rail, 2013f, p33 and 2014a, pp11-29).

| Section | Headway (minutes) | Sample Speed Limits (mph) |
|---|---|---|
| Standard Headway | 4 | |
| Exceptions: | | |
| Kings Cross to Finsbury Park | 3 (Fast Line) | 80 |
| | 4 (Slow Line) | 55 |
| Finsbury Park to Digswell | 3 (Fast Line) | 115 |
| | 4 (Slow Line) | 75 |
| Digswell to Woolmer Green | 3 | 115 |
| Woolmer Green to Hitchin | 3 (Fast Line) | 125 |
| | 4 (Slow Line) | 75 |
| Fletton to Peterborough | 4 (Fast Line) | 105 |
| | 5 (Slow Line) | 70 |
| Helpston to Stoke Junction | 4 (Fast Line) | 125 |
| | 5 (Slow Line) | 80 |

It can also be seen that the small range in headway values is in spite of a large variation in the speed limits of the various route sections. Furthermore, the Slow lines which are used by freight and ‘local’ stopping passenger services have lower sample speed limits than the associated Fast lines which are primarily intended for non-stop fast passenger services.

Headway values such as these make it is possible to calculate the maximum number of trains in a given time period using Equation (1).

$$C = \frac{T}{H} \tag{1}$$

Where:-

$C$ is the capacity (or maximum number of trains).

$T$ is the time period.

$H$ is the relevant headway.

Applying the values in Table 3.1 to Equation (1) gives a maximum capacity of between 20 trains an hour and 12 trains per hour (for 3 and 5 minute headways respectively). The headways of trains therefore have a significant influence on the possible capacity of rail infrastructure.

In this thesis, planning rather than technical headways will be used. This is because :-

- As noted they, unlike technical headways, are readily accessible.
- The calculations are significantly easier. For example, there is a close match between the sections used for the performance data and that used for the planning headways.
- The use of planning headways is consistent with previous work on capacity utilisation in Britain (e.g. Arup, 2013, p13).
- The concept of ‘planned’ or timetabled capacity utilisation is more relevant to the idea of the use of incentives through congestion charging than the use of the actual technical capacity utilisation.

### 3.3.7 The Calculation of Traffic Intensity

If the numbers of trains are known, then Equation (1) can be developed to calculate capacity utilisation as a percentage for that particular stretch of track. This is expressed as Equation (2).

$$I = \frac{N}{\left(\frac{T}{H}\right)} \times 100 \qquad (2)$$

Where:-

$I$ is Traffic Intensity (%)

$N$ is the number of trains in the given time period.

Equation (2) or ‘Traffic Intensity’ is the first one that will be used in this thesis to examine the relationship between capacity utilisation and performance. It can be seen that it is a function of train numbers and headway.

### 3.3.8 Timetable 'Compression' Methods

However, although it is expected that Equation (2) would be entirely effective when all traffic has the same characteristics (e.g. on dedicated 'High-Speed' lines) as noted by UIC (2004) the degree of heterogeneity is also an important factor in determining the capacity utilisation of a railway line. Indeed the creation of separate 'Fast' and 'Slow' lines, such as in the example given in Table 3.1, is intended to reduce the heterogeneity. Nash (1982) estimates that the provision of double-track line can as much as quadruple the potential overall capacity.

The impact of heterogeneity is illustrated in Figure 3.2. This shows a very simple 'time-distance' graph. Watson (2008) amongst others notes that this is a common approach for producing and expressing railway timetables.

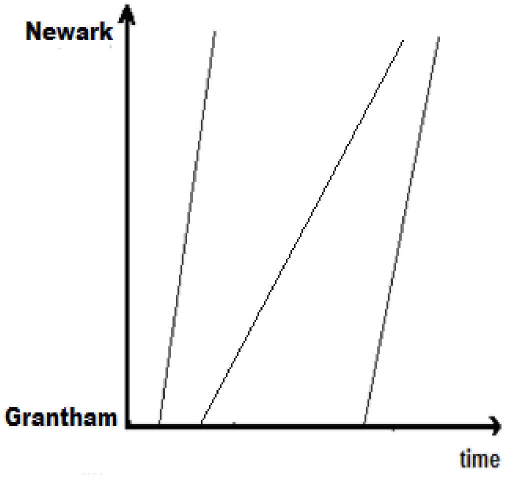


**Figure 3.2** Example of the Impact of Heterogeneity on Timetable Capacity.

In Figure 3.2 the example timetable between Grantham and Newark consists of three trains. The time taken to travel between the two locations is reflected by the slope of each of the train's lines, with the two faster trains having much steeper lines than the slower (central) train. It can be seen that the slower train clearly occupies more 'space' on the graph than the two fast trains and that by Newark it is beginning to be caught by the following 'fast' train.

This concept has been used as the basis for two popular methods of calculating capacity utilisation. These are the methods proposed by UIC (2004) which is widely used in mainland Europe (for example Schittenhelm and Landex, 2013); and the Capacity Utilisation Index (or CUI) approach which is widely used in Britain (for example Armstrong, J., Blainey, S., Preston, J. and Hood, I., 2011). Both ‘compress’ the trains in a timetable until they are the minimum headway apart. However, whilst the UIC approach uses technical headways the CUI approach uses planning headways. Therefore, for the reasons stated earlier in this chapter, the CUI approach has been adopted for this thesis.

Figure 3.3. applies the CUI ‘compression’ methodology to the example timetable seen in Figure 3.2.

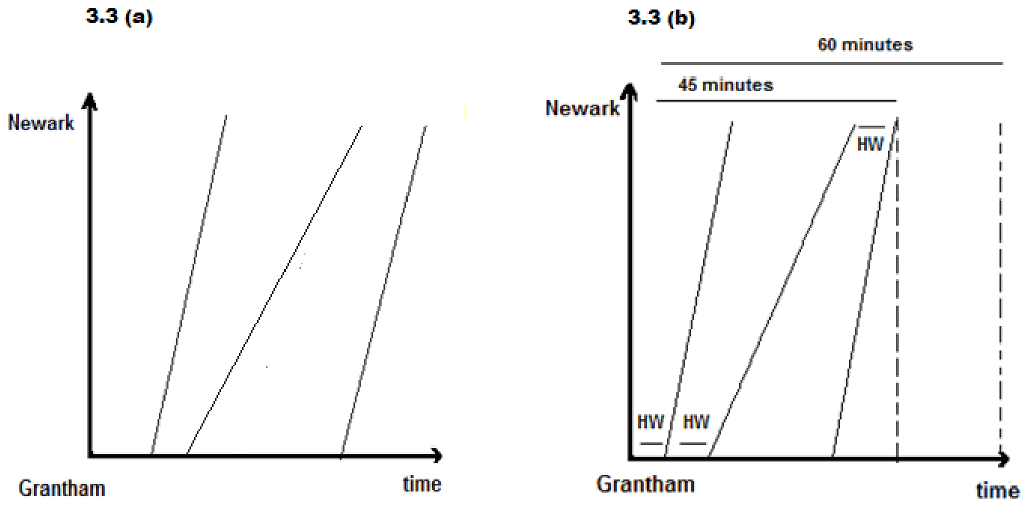


**Figure 3.3** Application of the CUI ‘Compression’ Methodology (source: Haith, J., Johnson, D., and Nash, C., 2014, p23).

Figure 3.3(a) shows the original non-compressed timetable with the second train in the sequence appreciably slower than the other two trains. The compressed state is shown in Figure 3.3(b). As noted by Gibson, S. Cooper, G., and Ball, B. (2002, p345) the CUI value equals the time occupied by the ‘compressed’ timetable divided by the time period.

This produces Equation (3).

$$OCUI = \frac{A}{T} \times 100 \qquad (3)$$

Where:

*OCUI* is the % capacity utilisation using the ‘original’ CUI method.

$A$ is the time period occupied by the compressed timetable.

$T$ is the original time period.

Using the values given in Figure 3.3 would therefore produce an OCUI value of 75%. Unlike, Traffic Intensity, OCUI takes into account the greater capacity utilisation of a heterogeneous timetable. Equation (3) will be used in the analysis carried out in this thesis.

Despite the popularity of the UIC / CUI “compression” approach in Europe, one of its problems is that it is currently largely confined to the calculation of ‘link’ only capacity utilisation. The UIC method recommends that “the line section used for compression should be reduced to the line section between two neighbouring stations (without overtaking or crossing possibilities)” (UIC, 2004, p18). Armstrong, J., Preston, J., Potts, C., Bektas, T. and Paraskevopoulos, D. (2013) note that the UIC themselves in a 2009 review of projects accepted that nodal capacity utilisation (i.e. at stations and junctions) has been largely ignored . This is probably because of its complexity. However, they also observe that there is limited value in calculating the capacity utilisation on the approach links to a station; if it is the platform occupancy within the station itself that is the main capacity constraint.

The added complexity of nodal capacity utilisation can be illustrated by considering the example of Newark Flat Crossing Junction on the East Coast Main Line. This is where the branch-line between Nottingham and Lincoln crosses the ECML just north of Newark North Gate station. The simple layout is shown in Figure 3.4.

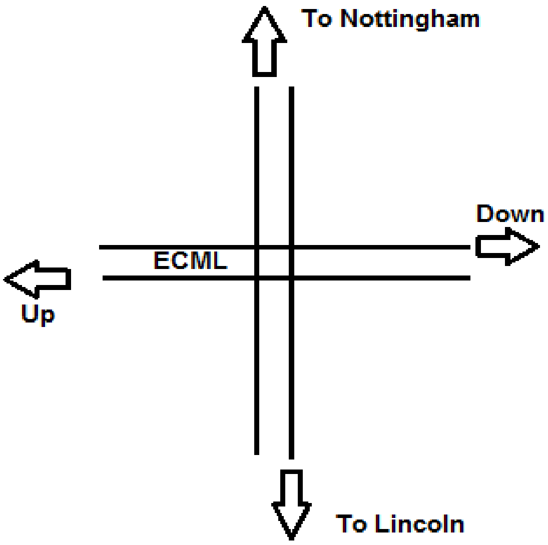
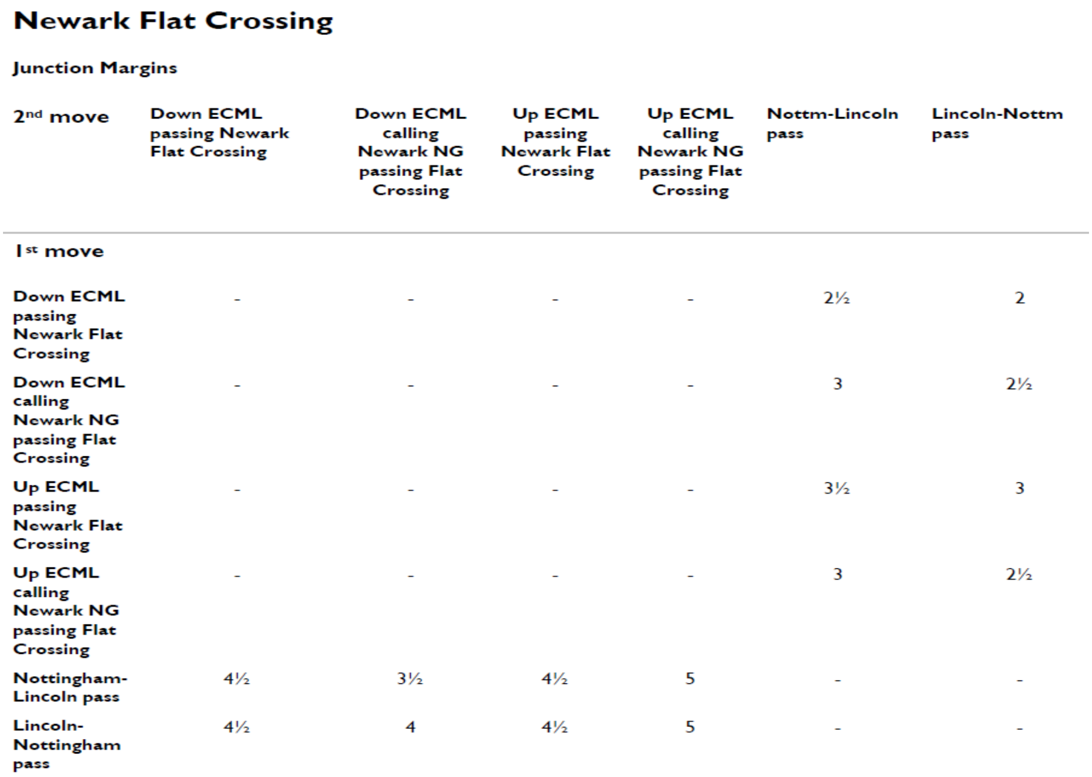


**Figure 3.4** Schematic Layout of Newark Flat Crossing (Based on Track Maps track diagrams, 2005, p16).

There are a total of four possible movements through the junction (ECML Up, ECML Down, towards Nottingham, towards Lincoln). Although there is no interaction between Up and Down main line traffic, there is between branch line and main line traffic. This is because the points have to be reset every time there is a ‘conflicting’ move.  The junction margin is the time between one move across the junction and the next conflicting move being allowed by the signalling system.

Once again there are planning margins and technical margins. For complex junctions (i.e. where there are a number of possibilities) the planning margins are presented in Network Rail’s Timetable Planning Rules in the form of a matrix. Figure 3.5 show those for Newark Flat Crossing.



**Newark Flat Crossing**

**Junction Margins**

| 2nd move | Down ECML passing Newark Flat Crossing | Down ECML calling Newark NG passing Flat Crossing | Up ECML passing Newark Flat Crossing | Up ECML calling Newark NG passing Flat Crossing | Nottm-Lincoln pass | Lincoln-Nottm pass |
|---|---|---|---|---|---|---|
| 1st move | | | | | | |
| Down ECML passing Newark Flat Crossing | - | - | - | - | 2½ | 2 |
| Down ECML calling Newark NG passing Flat Crossing | - | - | - | - | 3 | 2½ |
| Up ECML passing Newark Flat Crossing | - | - | - | - | 3½ | 3 |
| Up ECML calling Newark NG passing Flat Crossing | - | - | - | - | 3 | 2½ |
| Nottingham-Lincoln pass | 4½ | 3½ | 4½ | 5 | - | - |
| Lincoln-Nottingham pass | 4½ | 4 | 4½ | 5 | - | - |

**Figure 3.5** Junction Margins for Newark Flat Crossing (source: Network Rail, 2013f, p58).

Whilst the planning headways through the junction for East Coast Main Line traffic are simply 4 minutes in each direction, the junction margins are more complex. The values vary between 2 minutes and 5 minutes depending on the two ‘conflicting’ moves concerned. These are also affected by whether the main line traffic stops at the adjacent Newark North Gate station. It is clear that the exact pattern of traffic through the junction will have a substantial impact on the capacity consumption.

The margins in Figure 3.5 also suggest that capacity consumption at the junction itself will differ from the surrounding links. For example, as noted the minimum gap between successive ECML trains is four minutes. However, if the two trains (assuming they are Up trains and neither stop at Newark North Gate) have a Nottingham to Lincoln crossing the junction between them, then the gap between them at Newark Flat Crossing has to be eight minutes (i.e. 3.5 minutes between the first pair of moves and 4.5 minutes between the second pair of moves). The minimum capacity consumption would therefore be double at the node than for the adjacent link.

The approach adopted for calculating the capacity utilisation of links and nodes has been to develop a 'combined' compression approach. This is illustrated in Figure 3.6. Although, others (such as Armstrong, J., Preston, J., Potts, C., Bektas, T. and Paraskevopoulos, D., 2013) have measured junction nodes in isolation; the combined approach developed for this thesis assumes that there is a relationship between capacity utilisation at a node and on the adjacent links.

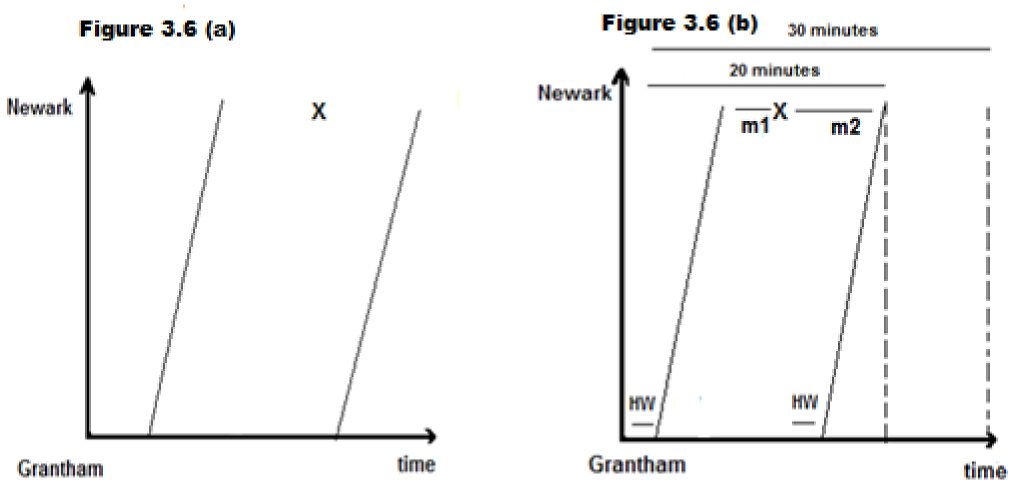


**Figure 3.6** The Inclusion of Crossing Moves in CUI Calculations Using the Combined Approach.

It can be seen that one of the three trains in the timetable used in Figure 3.2 has been replaced by a crossing move (denoted by an 'X'). In the compressed timetable shown as Figure 3.6 (b) it can be seen that the margin before the crossing move (denoted as 'm1' ) and after the crossing move (denoted as 'm2') dictate how far apart the two through trains are in the compressed timetable and consequently how much capacity is utilised. It can be seen that 20 minutes of the 30 minute time period are consumed giving a CUI value of 66.7%.

This produces Equation (4).

$$XCUI = \frac{AC}{T} \times 100 \qquad (4)$$

Where:

XCUI is the CUI value (%) for the combined (i.e. including links and nodes) ‘compressed’ timetable.

AC is the time period occupied by the combined compressed timetable.

$T$ is the original time period.

Equation (4) is used in the analysis carried out for this thesis.

It is also recognised that capacity at stations is also a factor in the potential capacity of a route. Station capacity is limited by the number of platforms they have and the layout of the tracks that access them. Stopping trains will consume platform capacity in the form of dwell times whilst passengers get on and off. Stopping trains also utilise additional track capacity due to the need to decelerate and accelerate. The rules governing station use are also documented in the Timetable Planning Rules and these could be used to calculate station capacity utilisation using the CUI method. Armstrong, J., Preston, J., Potts, C., Bektas, T., and Paraskevopoulos, D. (2013) examine this for a simple station layout. However, station capacity utilisation has been excluded from this analysis as it adds an extra degree of complexity. In the sample network used, adjacent links end with trains arriving at a station and begin with trains departing from them.

### 3.3.9 An Alternative Approach to ‘Compression’

It will be seen that the adoption of CUI as a metric provides consistency with previous work on the relationship between capacity utilisation and performance in Britain. However, as noted in Chapter One an objective of this thesis is to consider whether alternate philosophies are better able to explain this relationship.

A possibility is the work carried out by Vromans, M.J.C.M., Dekker, R. and Kroon, L.G.  (2006) on the measurement of heterogeneity on the Dutch rail network. They suggest a radically different approach to the timetable compression method. Whilst the latter measured the time trains ‘occupied’ a route section; Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006) recommended measuring the size of the actual ‘gaps’ between trains. This was on the basis that a train closely following the one in front would be more susceptible to delays.  The size of gaps will be a function of train numbers as well as heterogeneity. In the case of the former, gap size will decrease as the number of evenly spaced identical trains in a given time period increases. In the case of heterogeneity, the size of gaps will reduce as 'fast' trains begin to 'catch' 'slow' trains towards the end of route sections.

Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006) recognised that it was not appropriate to simply calculate the total sum of the gaps in an hour (because for example 20 trains multiplied by 3 minute gaps and 10 trains multiplied by 6 minute gaps both equal 60).  Instead they suggest the use of reciprocals. This has the advantage that smaller gaps have an increased weighting. In the previous example the reciprocal of 3 is 0.333 that of 6 is 0.167, giving a total of 6.66 compared to 1.67. Double the number of trains therefore produces four times the sum of the reciprocals of the gaps. Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006, pp653-654) use this approach in two equations.

Firstly, the 'Sum of the Shortest Headway[9] Reciprocal' (or SSHR) measures the point at which trains are closest to each other. Equation (5) gives the formula for calculating the SSHR:-

$$SSHR = \sum_{i=1}^{n} \frac{1}{H_i^-}$$

Where: (5)

$H^-$ is the smallest scheduled headway between train *i* and *i* +1.

This suggests that the gap following a train is the one used. For the analysis described in this thesis it was  assumed that the gap to measure was the one preceding a train (i.e. *i* and *i* -1) as it seems more logical that it is the size of this gap which will determine whether a train suffers delays.

Secondly, the conclusion was reached that the arrival gap at the end of a section should be used on the basis that "delays on arrival are on average more than delays at departure" (Vromans, 2005, p119). This observation led to creation of the 'Sum of the Arrival Headways Reciprocals' (or SAHR). There is however one potential issue with SAHR compared with SSHR.

[9] Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006) refer to scheduled headways. To avoid confusion with planning and technical headways these are referred to in this thesis as gaps.

Whilst it recognises the impact of a ‘fast’ train catching a ‘slow’ train towards the end of the section; it does not recognise the opposite situation of a ‘slow’ train following close behind a ‘fast’ train at the start of a section. SSHR measures the minimum gap wherever that might be. The effectiveness of both approaches will be compared as part of this thesis. The SAHR Equation is obviously similar to Equation (5) except $\mathrm{H}^{\mathrm{A}}$ is the arrival gap between two successive trains.

$$SAHR = \sum_{i=1}^{n} \frac{1}{H_i^A}$$

Where: (6)

$H^A$ is the arrival gap between train *i* and *i* +1.

One issue with both approaches is that neither compares the gaps with the minimum achievable gaps (i.e. the planning or technical headway). There is therefore no sense of how close to the maximum utilisation the timetable is (a timetable with trains five minutes apart will be considered ‘full’ if the planning headway is also five minutes but not if it is three minutes). The answer of course is to introduce these headways into the equation. This leads to the calculation of the ‘buffer’ time between trains.

Vromans (2005, p121) in his earlier Ph.D thesis recognised this short-coming and suggested calculating the ‘Adjusted Sum of Shortest Buffer Reciprocals’ (or ASSBR) using the following formula:

$$\mathrm{ASSBR(Q)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{1}{(\mathrm{H_i}^{-} - \mathrm{Hmin}_{\mathrm{i}}^{-}) + Q}$$

(7)

Where:-

$(\mathrm{H}^{-}– \mathrm{Hmin}^{-})$ is the smallest gap minus the planning headway (i.e. the ‘buffer’).

Q is described as the ‘average minimal headway’ or gap.

However, if the minimal headway Hmin is the same as H then the inclusion of Q leaves the SSHR described in equation (5). Excluding q would make the denominator 0. Indeed, Vromans (2005) himself acknowledges there are issues with this approach and it is noticeable that the ASSBR does not appear in the later paper by  Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006).

In preference, a new way of including the planning headway has been devised [10]. This is shown in Equation (8).

$$\text{OHET} = \frac{SSHR}{\frac{1}{PH} \text{ x } G} \text{ x } 100. \qquad (8)$$

Where:

OHET is the calculated OHET in % .

SSHR is derived using formula (5)

PH is the planning headway.

G are the number of ‘gaps’ in the timetable.

Equation (8) compares the reciprocal of the observed ‘gap’ with the reciprocal of the headway. For example, a train 9 minutes behind the train in front (giving a reciprocal of 0.111) on a section with a 3 minute headway (giving a reciprocal of 0.333) would have a calculated OHET value of 33.3%. If the train was the minimum headway behind the train in front the OHET value would be 100%.

At this point it is interesting to compare OCUI values that can be obtained with some OHET values that assume two different types of spacing within a one hour time period. These are presented as Example One which is shown in Figure 3.7.

**Example One**

5 trains in an hour with identical characteristics and 3 minute planning headway.

**OCUI calculation – Equation (3)**

Compressed timetable occupies 15 minutes (5 x 3). 15 /60 * 100 = **25%**

**OHET calculations – Equation (8)**

Scenario 1 – Even Spacing i.e. every 12 minutes

Reciprocal of 12 minute gaps = 0.083 (Total for five trains = 0.415)

[10] First described in Haith, J., Johnson, D. and Nash, C.  (2014)

Reciprocal of 3 minute headway = 0.333 (Total for five trains = 1.665)

**Example One Continued**

0.415 / 1.665 *100 = **25%**

Scenario 2 – Irregular Spacing (5 + 15 + 8 + 11 +21 = 60 minutes)

Reciprocals = 0.200 + 0.067 + 0.125 + 0.091 + 0.048 = 0.581.

Reciprocal of 3 minute headway = 0.333 (Total for five trains = 1.665)

0.531 / 1.665 * 100 = **31.9%**

**Figure 3.7** Example One - Sample OCUI and OHET calculations.

Figure 3.7 shows an example OCUI calculation with two OHET calculations using the same number of trains and planning headway. The HET scenarios assume two different approaches to timetable spacing . In the example, the trains have identical characteristics. Whilst heterogeneity due to different characteristics is 'captured' by both approaches, albeit in different ways, it is interesting to observe the impact of using identical trains. It can be seen that the OHET percentage is the same as the OCUI percentage when the timetable is evenly spaced. However, irregular spacing increases the OHET percentage, whilst OCUI which does not take into account actual timetable spacing remains the same. This suggests that CUI assumes even spaced timetables when the train characteristics are identical, and also serves to validate the HET approach. It will be seen that the impact of irregular spacing or 'bunching' on timetable performance is a key theme of this thesis.

OHET i.e. Equation (8) is used in the analysis carried out for this thesis. Arrival HET or AHET substitutes SAHR for SSHR in Equation (8) to give Equation (9). Equation (9) can be expressed as follows:

$$\mathrm{AHET} = \frac{SAHR}{\frac{1}{PH} \text{ x } G} \text{ x } 100. \qquad (9)$$

Where:

AHET is the calculated AHET in %.

SAHR is calculated using equation (6) and uses the arrival gaps rather than the minimum gaps.

PH is the planning headway.

G are the number of 'gaps' in the timetable.

This is also used in the analysis carried out for this thesis. The next step was to develop Equation (8) to take into account the impact of junction crossing moves. Trains are also likely to be effected by how close in front of them a crossing move is planned. As with the consideration of CUI for links and nodes, a combined approach was produced. It will be seen that an equal weighting approach was applied to the headways and margins. In reality the actual weighting could vary from node to node. However, taking the mean of the link and junction gaps keeps the approach as simple as possible. To calculate the combined capacity utilisation of links and nodes, Equation (10) was devised.

$$\text{XHET} = \frac{\text{SSHR} + \sum_{\text{i}=1}^{\text{n}} \left( \frac{\frac{1}{\text{LXG}_\text{i}} + \frac{1}{\text{AJXG}_\text{i}}}{2} \right)}{\frac{1}{\text{PH}} \times \text{G}} \text{ X } 100 \qquad (10)$$

Where:-

$SSHR$ is the sum of Sum of Shortest Headway Reciprocals for trains without a crossing move planned in-front of them.

$n$ is the number of trains with a crossing move planned in-front of them.

$G$ is the number of gaps in-front of trains.

$LXG$ is the observed gap to the previous ‘through’ train for a train that has a crossing move planned in-front of it.

$AJXG$ is the adjusted observed gap to a previous crossing train.

It is necessary to provide some explanation of the adjustment process (used to produce $AJXG$). This is where the size of the gap is adjusted so that the headway can be used as the denominator for all gaps. For example, for an observed junction crossing gap of 8 minutes with a junction margin of 5 minutes but a headway of three minutes (i.e. the denominators differ by two minutes) then the observed gap is adjusted by minus 2 minutes to give 6 minutes  The size of the ‘buffer’ (i.e. 3 minutes greater than the margin or headway) is therefore retained but consistency is  maintained within the calculations. In order to further clarify how XHET is calculated a more detailed example is presented in Figure 3.8.

**Example Two**

3 ‘Southbound’ through trains with identical characteristics in the time period.

3 Crossing moves (2 in one direction, 1 in the other).

3 minute Headways.

2 minute Junction Margin (first train through train / second train crossing move – **not taken into account** in the calculation as only the impact on through trains is of interest).

5 minute Junction Margin (first train crossing move / second train through train – **this is of interest**).

It can be seen that Train A does not have a crossing move in-front of it. In contrast both trains B and C do.

The gap from Train a to Train A is taken into account, even though Train a is planned in the previous time period.

The gap from Train 1 to Train B is not taken into account as Train 2 has a smaller gap.

**The Timetable**

| | Train *a* | Train A | X Train 1 | X Train 2 | Train B | X Train 3 | Train C |
|---|---|---|---|---|---|---|---|
| Timing Point A | ww40 | xx00 | | | xx20 | | xx40 |
| **Junction** Timing Point B | ww50 | xx10 | | | xx30 | | xx50 |
| West to East | | | xx22 | | | xx39 | |
| East to West | | | | xx23 | | | |

**Example Two Continued**

Train A gap is 20 minutes to Train *a* giving a reciprocal of **0.05**.

Train B gaps are 20 minutes to Train A (giving a reciprocal of 0.05) and 7 minutes to Train 2 adjusted to 5 minutes (as 7-5 = 5-3). This gives a reciprocal of 0.2. The average of 0.05 and 0.2 is **0.125**.

Train C gaps are 20 minutes to Train B (giving a reciprocal of 0.05) and 11 minutes to Train 3 adjusted to 9 minutes (as 11-5=9-3). This gives a reciprocal of 0.111. The average of 0.05 and 0.111 is **0.161**.

The sum of the reciprocals for the gaps is 0.05 + 0.125 + 0.161 = 0.272 (i.e. the SSHR)

The reciprocal of the headway is 0.333 (3 trains giving a total of **0.999**) (i.e. the HW).

0.272 / 0.999 gives a XHET value of **27.2%**.
The associated link-only HET calculation would produce a value of **15%**.
(0.05 x 3 = 0.15 / 0.999 = 0.15)

**Figure 3.8** Example Two – Sample XHET Calculation

It can be seen that the inclusion of junction moves in Figure 3.8 almost doubles the calculated capacity utilisation. Equation (10) is used in the analysis carried out in this thesis.

Up until this point, the assumption has been made that all ‘gaps’ are of equal importance. However, Carey (1999) suggested that it might be advantageous to weight the size of gaps for different types of train when investigating the relationship between capacity utilisation and timetable performance. For example, in a largely regular interval passenger timetable with several long distance freight trains it might be beneficial to give the freight trains larger buffers. This introduces the idea of ‘vulnerable’ trains which are either more likely to cause reactionary delay or be susceptible to it. To test the validity of this, Equations (11) and (12) have been developed. Equation (11) gives added weight to the ‘buffer’ preceding a vulnerable train by effectively counting it twice in the calculation.

$$\text{VHETB} = \frac{SSHR{\sim} + \sum_{i=1}^{V}(2VGB_i)}{\frac{1}{PH} \times (G+V)} \text{ X } 100 \qquad (11)$$

Where:

$SSHR\sim$ is the SSHR for all non-vulnerable trains in the time period.

V is the number of vulnerable trains in the time period.

$VGB$ is the ‘gap’ preceeding a ‘vulnerable’ train in the time period.

$G$ is the number of gaps in the timetable.

Equation (12) substitutes VGF (the gap following a ‘vulnerable’ train in the time period) for VGB. This is based on the concept that a following train has a greater risk of delay due to the vulnerable train having an assumed increased chance of performance problems. As with the weighting applied to XHET, the weightings for the two VHET equations are rather arbitrary and would therefore benefit from future study. The purpose of including them here is simply to investigate the possibility that the concept is sound.

Equation (14) can be expressed as follows:-

$$\mathrm{VHETF} = \frac{SSHR\sim + \sum_{i=1}^{V}(2VGF_i)}{\frac{1}{PH} \times (G+V)} \text{ X } 100 \tag{12}$$

Where:

$SSHR\sim$ is the SSHR for all non-vulnerable trains in the time period.

V is the number of vulnerable trains in the time period.

$VGF$ is the ‘gap’ following a ‘vulnerable’ train in the time period.

$G$ is the number of gaps in the timetable

Both Equations (11) and (12) are used in the analysis carried out for this thesis.

### 3.3.10 The Influence of Critical Links and Nodes

One other aspect that needs to be considered, when discussing the measurement of capacity utilisation, is the idea of critical links and nodes. Researchers including Kraft (1982) and Burdett and Kozan (2006) have suggested that the potential capacity of a network can be determined by measuring the capacity utilisation at its most constrained points. Furthermore, as described at the start of this chapter, ‘Bottle-neck’ congestion is one of Vickrey's (1961) six classifications of general traffic congestion.

This concept is not restricted to rail, or even the wider transport field. For example, Goldratt in the early 1980s (Goldratt and Cox, 2004) popularised the idea with his business improvement methodology 'The Theory of Constraints'. This was first applied to production line manufacturing where he noted that the output of the entire line was dictated by the speed of the slowest machine or process. By improving the flow through this constraint, the output of the entire line can be increased. Since then the Theory of Constraints has been applied to a wide variety of disciplines. Mabin and Balderstone (2000) in their book recording the use of the technique, for example, mention improving Health Service provision and Software design amongst the more obvious manufacturing applications.

The Theory of Constraints therefore appears to be a philosophy that could be valuable in furthering our understanding of the relationship between capacity utilisation and performance on congested rail networks. In order to test this possibility two new measures LCUI and LHET, standing for Local CUI and Local HET respectively, were devised. These use the XCUI and XHET equations (Equations (4 ) and (10) respectively) but are solely calculated for the identified primary constraints in a sample network. The relationship between the values and the performance of the surrounding rail network is then investigated.

As a final further test EHET, or Expanded HET, was devised. This examines the relationship between the minimum timetabled gap in a sample rail network, wherever that may be with the overall performance. The purpose behind this measure, which uses Equation (8) as a base, is to establish whether timetabled gaps at the primary constraints or the overall minimum gaps are the most important.

### 3.3.11 Summary of the Capacity Utilisation Measures used in this Thesis

Table 3.2 summarises the capacity utilisation measures used as explanatory variables in the analysis carried out for this thesis.

It can be seen that there are three basic approaches (I, CUI and HET) with the latter two being further sub-divided. As noted earlier in the chapter, Traffic Intensity (I) only takes into account train numbers whilst CUI and HET account for heterogeneity as well. The latter two therefore more closely reflect the Capacity Balance in Figure 3.1 and as such are expected to be more accurate measures of capacity utilisation.

**Table 3.2** Summary of the Capacity Utilisation Measures Used in the Analysis Carried out for this Thesis.

| Abbreviation | Name | Type | Equation |
|---|---|---|---|
| I | Traffic Intensity | Number of Trains | (2) |
| OCUI | Capacity Utilisation Index | Link occupation | (3) |
| XCUI | Capacity Utilisation Index (including crossing moves) | Link and Junction occupation | (4) |
| OHET | Heterogeneity Measure | Link ‘Buffer’ Times | (8) |
| AHET | Arrival Heterogeneity | Link Arrival ‘Buffer’ Times | (9) |
| XHET | Heterogeneity Measure (including crossing moves) | Link and Junction ‘Buffer’ Times | (10) |
| VHETB | Heterogeneity Measure including Vulnerable trains element | Link Adjusted ‘Buffer’ Before Times | (11) |
| VHETF | Heterogeneity Measure including Vulnerable trains element | Link Adjusted ‘Buffer’ Following Times | (12) |
| LCUI | Local CUI Measure | Constraint Occupation | Uses (4) |
| LHET | Local HET Measure | Constraint ‘Buffers’ | Uses (10) |
| EHET | Expanded HET Measure | Area Minimum ‘Buffers’ | Uses (8) |

It is also worth reinforcing the point that whilst HET measures ‘buffer’ times between individual trains these are ‘lost’ in the CUI calculations as the timetable is compressed to the minimum headways and margins. Furthermore, CUI assumes that trains with identical characteristics will be evenly spaced in the timetable whilst HET distinguishes between the spacing of identical trains.

## 3.4 Timetable Performance

### 3.4.1 Some Basic Definitions

In a similar way to capacity utilisation, the definition of rail performance varies depending on the use for which the information is intended. In Britain, rail performance is measured in two distinct ways: those that have a public purpose and those that are for internal use within the rail industry. In both cases ‘delivery’ on the day is compared with the ‘plan’ contained in the timetable. Divergence from the plan will be a result of trains arriving at their destinations later than advertised, suffering delays en-route or being cancelled in their entirety.

Whilst ‘delay’ can be defined as minutes lost between two consecutive timing points; ‘lateness’ is the overall timing of a delayed service at a certain point in its journey (in performance terms usually its destination) and thus reflects the cumulative impact of delays and any recovery allowances en-route.

The Public Performance Measure (or PPM) is the one that is shared with the general public and is used as a target by the ORR for the Train Operators and Network Rail. As described by Network Rail (2015), PPM is divided into ‘Punctuality’ and ‘Reliability’. Punctuality expresses the percentage of services arriving at their destination ‘on-time’. Services are counted on-time if they arrive at their destination within a certain threshold of their scheduled arrival time (this is within 10 minutes for long-distance services and 5 minutes for all other services). ‘Reliability’ reflects the level of cancellations, with any cancelled or part-cancelled (i.e. at an intermediate point on the journey) being counted as ‘late’ for the purposes of the measure. Together the measure expresses the percentage ‘on-time’ arrival for trains. PPM will be returned to later in this thesis as it does have some relevance to the subject of this thesis. However, one of its shortcomings for comparison with capacity utilisation is that it does not take into account performance en-route.

In contrast to the PPM, timetable delays provide information on performance en-route. This means that they can be more closely linked to the capacity utilisation measures described earlier in this Chapter.  In Britain, delay minutes are divided at the attribution stage into Primary and Reactionary Delays (Delay Attribution Board, 2011) (which are commonly called Secondary delays in other countries). Primary Delay is the direct delay caused to train services by a performance incident. For example, trains that have to stop whilst a failed set of points is repaired would have their delay counted as primary delay. In contrast, Reactionary Delay is indirect delay to train services that arise due to the incident. This is where additional delay is caused to services as a result of trains running late following the performance incident. For example, if one of the late running trains due to the points failure then itself caused a train elsewhere on the network to be delayed, the delay to the second train would be counted as reactionary delay. A subset of reactionary delay is Congestion Related Reactionary Delay (or CRRD). These are reactionary delays that at the attribution stage have been coded as being associated with congestion on the line.

**Table 3.3** CRRD codes used in TRUST (Source: Arup, 2013, p14)

| **Reactionary Delay Code** | **Description** |
|---|---|
| YA | Lost Path : Regulated for Train running on Time |
| YB | Lost Path : Regulated for another late running Train |
| YC | Lost Path : Following Train Running on Time |
| YD | Lost Path : Following another later running Train |
| YE | Waiting Acceptance to Single Line |
| YF | Waiting for Late Running Train off Single Line |
| YG | Regulated for Late Running High Priority Train |
| YO | Waiting due to platform / station congestion or platform change |

Table 3.3 gives the CRRD codes that are used by Network Rail in their performance monitoring and attribution system TRUST[11].

CRRD is clearly a very useful measure of the impact of capacity utilisation on timetable performance. The level of CRRD is therefore the performance metric (i.e. the dependent variable) that will be used in this thesis to examine the relationship between capacity utilisation and performance. Its use also maintains consistency with previous work on the subject in Britain (Gibson. S., Cooper, G. and Ball, B., 2002; Faber-Maunsell, 2007 and Arup, 2013).

### 3.4.2 Previous Relevant Research

There has been a great deal of academic interest in the relationship between capacity utilisation and the level of reactionary delays. Research has used a wide variety of techniques ranging from the correlation of observed data to the modelling of theoretical rail networks using various simulation techniques. The scale and complexity of the areas studied has also varied from short single lines to entire rail networks with many links and nodes.

The volume of traffic (or traffic intensity) has been identified as a key factor in the development of reactionary delays. This has been noted by Brunel, J., Marlot G. and Perez M. (2013); Lindfeldt A. (2012); Lindfeldt O. (2010); Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008); Goverde (2007); Higgins A., Kozan E., Ferreria L. (1995); Carey and Kwiecinski (1994) and Petersen (1974).

Higgins A., Kozan E., Ferreria L. (1995) found that adding an additional train to a theoretical single-line railway led to slightly more primary delay due to there being more trains that could be affected. However, the amount of reactionary delay increased substantially. Dingler, M.H., Lai, Y. and Barkan, C.P.L. (2009, p43) in a simulation of a hypothetical 124 mile long single-line in North America found that “the effect of additional trains on delay is not linear. Instead, the relationship between train volume and delay is exponential”. Sogin, S., Barkan, C. and Saat, M. (2011) found that for completely homogenous freight traffic on a single-line, delays were found to increase exponentially with traffic density. It should be noted though that these three studies were for single-lines which primarily or completely contained freight traffic.

[11] TRUST stands for Train Running System on TOPs with TOPS standing for Total Operations Processing System. (Railway-Technical.com, 2013).

Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008) used the term ‘congestion’ in their description of the theoretical relationship between increasing traffic volume and reactionary delays. Their diagram is reproduced as Figure 3.9. It can be seen that three levels of traffic density are defined: ‘normal’, ‘saturated’ and ‘congested’ and Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A (2008) explain how average delays increase dramatically and network reliability is rapidly lost once a congested level is reached.

Figure 3.9 suggests that the relationship between capacity utilisation and performance is an exponential one with a very steep upwards curve. The idea that the relationship is exponential is clearly an important observation and this will be returned to later in this thesis.

More traffic means a much greater susceptibly to reactionary delays which leads to more traffic receiving yellow or red aspects and an unstable stop and start relationship commences.

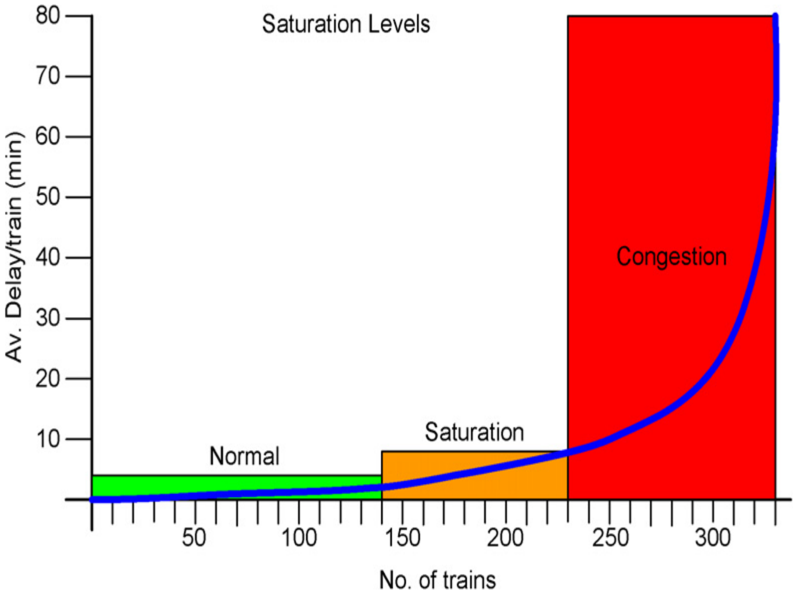


**Figure 3.9** The Relationship Between Traffic Volume and Average Delays (Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A., 2008, p780).

Lindfeldt, A. (2012); Sogin, S., Barkan, C. and Saat, M. . (2011); Lindfeldt O. (2010); Dingler, M.H., Lai, Y. and Barkan, C.P.L.. (2009); Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006) and Huisman and Boucherie (2001); have all identified a link between the degree of heterogeneity in a timetable and the level of reactionary delay. Vromans, M.J.C.M., Dekker, R. and Kroon, L.G..(2006, p647) in their work on the Dutch rail network note that “the shared use of the same infrastructure by different railway services, with different origins and destinations, different speeds, and different halting patterns, is probably the main reason for the propagation of delays throughout the network”. They found that calculated average delays were substantially higher for a heterogeneous timetable than a homogeneous timetable with the same number of trains.

There is considerable agreement why higher levels of traffic and heterogeneity tend to result in higher levels of reactionary delays. Landex (2008); Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. (2006); Huisman and Boucherie (2001) and Higgins A., Kozan E., Ferreria L. (1995); all refer to the existence of small buffer times between trains being the prime cause of increased levels of reactionary delay. Carey (1999) used a heuristic approach to link reactionary delay and the reliability of the timetable with the size of the ‘buffer’. He calculated that the probability of reactionary delays occurring was decreased by making the gaps between trains equal for those with the same characteristics. Yuan and Hansen (2007) found that as scheduled ‘buffer’ times between trains were reduced, reactionary delay increased at an exponential rate. Lindfeldt, A (2012) linked heterogeneity to the size of buffer times in two ways. Firstly he referred to an uneven distribution of trains in the timetable resulting in reduced buffer times. Secondly, he referred to a mix of traffic of different characteristics which increased the likelihood of ‘fast’ trains catching ‘slow’ trains.

However, there is an alternative view noted by Watson (2008, p126). He suggested that although on simple networks equal spacing should produce lower levels of reactionary delay, a different strategy could be more effective on more complex networks. This was the planning close together, or ’flighting’, of similar trains. This would mean that at junctions, for example, movements from the same flow would ‘clear’ the junction more quickly. Both points of view will be explored as part of this thesis.

Finally, a number of attempts have been made to define the point at which it is not sensible to add any more traffic to a network. Burdett and Kozan (2006) refer to the difference between absolute capacity and sustainable capacity.  UIC in their document on rail capacity (UIC, 2004)  have gone so far as to propose guideline values of capacity utilisation, derived using the compression methodology described earlier, beyond which the infrastructure should be declared congested and no more train paths accepted. These values are shown in Table 3.4. They note that these are based on current practice by European Infrastructure Managers.

The table shows different values for peak hours and the daily period. This highlights the fact that peak traffic levels can only be sustained for limited periods of the day if the network is going to be able to recover quickly from major performance incidents. The table also highlights differences in recommended capacity utilisation values for the different types of line. It can also be seen that utilisation can be increased further if certain conditions are met.

**Table 3.4** Recommended Maximum UIC Capacity Utilisation Values for Different Types of Line (UIC, 2004, p19).

| Type of Line | Maximum Peak Hour Utilisation | Maximum Daily Period Utilisation | Comments |
|---|---|---|---|
| Dedicated passenger suburban traffic | 85% | 70% | The possibility to cancel some services in case of delays allows for high levels of capacity utilisation |
| Dedicated high-speed line | 75% | 60% | |
| Mixed-traffic lines | 75% | 60% | Can be higher when number of trains is low (smaller than 5 per hour) with strong heterogeneity |

Previous research therefore suggests a strong link between the capacity utilisation of a timetable and its performance (as measured by the level of reactionary delay), with a key factor being the size of the buffer between successive trains.

### 3.4.3 Other Factors that Need to be Taken into Account

Three different types of capacity utilisation measurements have so far been identified (Traffic Intensity, CUI and HET) that will be used as possible Explanatory Variables in the analysis carried out for this thesis. Whilst Traffic Intensity only measures one element of UIC's Capacity Balance Diagram (Figure 3.2) both CUI and HET capture all four elements of traffic numbers, heterogeneity, stability and average speed. This is because each element affects how much time a compressed timetable occupies in the case of CUI and the size of the 'buffer' between individual trains in the case of HET. However, whilst the inclusion of traffic numbers and heterogeneity are intuitive in the CUI and HET calculations, the stability of the timetable and average speed deserve further attention.

Stability has been specifically linked by a number of researchers to the level of reactionary delays. Yuan and Hansen (2007) suggest that the level of reactionary delays reflect the robustness of a timetable and the stability of train operations. Goverde (2007) defines stability as the ability of the timetable to absorb delays so that they do not propagate.

The addition of extra time into train schedules to cope with minor unexpected delays (referred to as Performance Allowances in Britain) whilst increasing the stability of the timetable will increase the amount of time it occupies (or reduce the timetabled gaps between trains). Similarly, the addition of extra journey time into a train schedule to maintain the headway behind the train in-front (referred to as Pathing Allowances in Britain), because it requires the train in question to slow down, can also theoretically allow some 'recovery' of delays if the train is actually able to operate at its normal speed. In order to investigate the relationship between the amount of allowances in a timetable and the level of reactionary delay, Equation (13) was devised.

$$Stability = \frac{PET+PAT}{n} \tag{13}$$

Where:

$Stability$ equals the average allowances included in a given time period.

$PET$ is the total amount of Performance Allowances in minutes added.

$PAT$ is the total amount of Pathing Allowances in minutes added.

$n$ is the number of trains in the time period.

Landex (2008) suggested a link between complexity and stability. He suggested that complexity be measured as a function of the possible interactions between trains at junctions and stations. As noted by Landex his detailed approach is only concerned with infrastructure stability rather than the stability of operations which includes that of the timetable. For this reason his method has not been pursued as part of this thesis. The idea that there is a relationship between complexity of the timetable and the level of reactionary delays does however need to be pursued.

Network Rail in a Rail Industry Seminar held in May 2013 (Network Rail 2013d) expressed timetable complexity in terms of the number of Service Codes seen on a given route and time period. Service Codes sub-divide an operator's paths by their general route. For example, East Coast services on the East Coast Main Line have separate Service Codes for Leeds to London; Edinburgh, Glasgow and Newcastle to London and Aberdeen or Inverness to London.

Equation (16) is therefore simply:-

$$T\ T\ Complexity\ = No.SCs \tag{14}$$

Where:

$No.SCs$ equals the number of Service Codes in the time table for a given route section and time period.

It will be seen later in the thesis that the use of Service Codes as one of the factors considered is rather appropriate.

Average speed is perhaps more straight-forward to consider than timetable stability. Differences in speed will be reflected in both CUI and HET measures due to the increased level of heterogeneity. However, Gibson, S., Cooper, G. and Ball, B. (2002) examined separately the impact that speed differences outside the prevailing range on a route would have on delay. They did this through the use of simulation. Although, small differences in speed were found to not have a significant effect on delay, speeds that were substantially greater than the maximum or substantially lower than the minimum were found to produce significantly more delay. To investigate this, Equation (15) was devised. Due to data availability this uses the transit time in minutes for a given route section, rather than speed and is entitled Average Transit Time Variation .

$$\text{Av. Transit Time Variation} = \frac{\sum_{i=1}^{n}\left(\frac{AT_i}{BMT}\right)}{\text{n}} \tag{15}$$

Where:

$Av.Transit\ Time\ Variation$ is the average difference between actual and base transit time for a given time period and route section.

$n$ is the number of trains

$AT$ is the actual trainsit time for a given train. This includes any allowances.

$BMT$ is the Base or Minimum transit time for the route section in question.

A number of other causal factors in the development of reactionary delay have been identified which need to be considered. Kraft (1982) observed that a rail line can operate close to maximum capacity utilisation providing it has a low risk of primary delays. The risk of delays could be reduced for example through more frequent renewal, inspection and maintenance of infrastructure assets. However, this would add to the cost of the rail network. Ferreria (1997) makes the point that the type and availability of the rail infrastructure has an impact on the overall level of delays. In particular he referred to the importance of placing sidings and passing loops in the best positions on new single track lines if overall delays are to be minimised. Lindfeldt O. (2012) using Railsys simulation also found that infrastructure factors were significant but affected the level of reactionary delay in a complicated way. A detailed analysis of these factors is outside the scope of this thesis since the intention is to consider simple theoretical relationships. As described earlier, the work of Krueger (1999) for example, has examined the role of infrastructure factors in the level of capacity utilisation and the consequent levels of delay.

One aspect that will be pursued is the relationship between a given route section and the rest of the network. Goverde (2007) used an example from the Dutch rail network to show that reactionary delays would propagate widely in highly inter-connected timetables as well as in those with high traffic densities. A useful step in considering this is to examine the relationship between a route section and the surrounding network. The concept of ‘critical sections’ effecting the available capacity and performance of the surrounding network has already been discussed.

It is also logical that trains entering a route section which are already late-running will be more susceptible to reactionary delay. To examine this concept the average recorded lateness for each train entering the relevant route section in a given time period will be used. Although in his analysis of the causes of delay propagation, A.Lindfeldt (2012) using simulation of the

Swedish network did not find a significant relationship between ‘entry-lateness’ and reactionary delay the concept does seem to be worth further consideration. Equation (16) has therefore been developed to test this relationship:-

$$\text{Av. Entry Lateness} = \frac{\sum_{i=1}^{n}(ELT_i)}{\text{n}} \tag{16}$$

Where:

$Av.Entry\ Lateness$ is the average entry lateness for all trains for the relevant route section and time period.

$ELT$ is the average entry lateness over time for a specific train for the relevant route section and time period.

$n$ is the number of trains.

Olsson and Haugland (2004) referred to regression analysis of causal factors of punctuality in Britain in the 1990s. A significant factor was found to be average distance travelled. Although, referring to punctuality at final destination rather than the level of reactionary delay in a given geographic section, this would seem to be a factor worth investigating. It can be theorised that the further a train has travelled the more delays it will have already suffered en-route making reactionary delay in the given section more likely. It can also be surmised that long distance trains in Britain connect otherwise unrelated parts of the network increasing the likelihood of delay propagation. Therefore, in order to investigate the relationship between average distance travelled and the level of reactionary delay Equation (17) was used.

$$\text{Av. Distance Travelled} = \frac{\sum_{i=1}^{n}(DT_i)}{\text{n}} \tag{17}$$

Where:

$DT$ equals distance travelled from train origin to start of the relevant route section.

$n$ is the number of trains in the time period and the relevant route sections.

Finally, a related idea is the one that capacity utilisation in the surrounding network and the preceding time period will affect the performance of the route section in question. In order to investigate this three variants of Intensity, CUI and HET were created. These are ‘Time Before’, ‘Section Before’ and ‘Section Following’. They use the relevant capacity utilisation figure for the previous time period or adjacent link as appropriate. For example, the ‘Time Before’ ‘Intensity’ for the time period 0800 to 0900 would be the calculated ‘Intensity’ for the period 0700 to 0800. Example Three which is given in Figure 3.11 illustrates ‘Section Before’ and ‘Section Following’.

**Example Three**

Network with three sections AB, BC and CD.

Section Before (SB) for BC is the calculated capacity utilisation for AB.

So SBCUI for BC is the CUI value for AB.

Section Following (SF) for BC is the calculated capacity utilisation for CD.

So SFCUI for BC is the CUI value for CD.

**Figure 3.11** Example Three – The Calculation of the ‘Section Before’ and ‘Section Following’ Capacity Utilisation Measures.

### 3.4.4 Summary of Equations for the ‘Other’ Factors

**Table 3.5** Summary of the ‘Other’ measures used in this thesis.

| Abbreviation | Name | Equation |
|---|---|---|
| STAB | Stability | (13) |
| TTC | Timetable Complexity | (14) |
| ATV | Average Transit Time Variation | (15) |
| AEL | Average Entry Lateness | (16) |
| ADT | Average Distance Travelled | (17) |
| TBCAP | Capacity Utilisation for Time Period Before | As appropriate |
| SBCAP | Capacity Utilisation for Section Before | As appropriate |
| SFCAP | Capacity Utilisation for Section Following | As appropriate |

## 3.5 Summary

Using a literature review this chapter has explained the general principles of traffic congestion and discussed previous research into the relationship between capacity utilisation and performance. The measures used to represent capacity utilisation have been discussed in some detail. The creation of these explanatory variables has been explained together with explanatory variables for the ‘other’ possible causes of reactionary delay that will be investigated. Congestion Related Reactionary Delay (or CRRD) has been identified as the metric (or dependent variable) to be used in this thesis to represent performance.

The explanatory variables and dependent variable described in this chapter will be investigated using the methodology outlined in Chapter 5 and the data set described in Chapter 6.

Two important observations on the relationship between capacity utilisation and performance have been made. Firstly, it has been suggested that the relationship is exponential in nature. Secondly, a number of researchers suggest a prime cause of reactionary delay is the size of gaps (or buffers) between trains.

A number of other measures have also been identified as either complimentary or different causal factors for observed levels of reactionary delay.

The next chapter builds on the literature review by discussing the actual nature of performance on Britain's rail network and the causal factors identified by the industry itself. The development of Britain's Capacity Charge is then explained in some detail.

# Chapter 4
# Actual Timetable Performance and the Development of the Capacity Charge in Britain

## 4.1 Introduction

This chapter examines the issues affecting the actual performance of the rail network in Britain as identified by the industry itself. This is compared with the findings of the literature review described in Chapter Three. The privatised nature of the Britain's rail network leads to a considerable amount of performance monitoring and investigation. This means that there is a substantial amount of data and information available.

The second part of this chapter explores the creation and development of the Capacity Charge in Britain. This 'congestion' charge was first introduced in 2002. The recent recalibration exercise carried out in 2013 produces an excellent framework for the analysis carried out as part of this thesis.

## 4.2 Rail Performance in Britain

### 4.2.1 Overview

As referred to earlier in this thesis, the rail industry uses a system called TRUST. This database compares the working timetable, or train plan, with the actual recorded time at the key locations for individual services on a given day and is used to record the cause of any delays and the organisation deemed responsible for it (Delay Attribution Board, 2011).

In Britain, performance is measured intensively. Vertical separation of the industry following privatisation and the introduction of track access charges meant that there needed to be a way for the infrastructure owner and the train operators to understand whether the purchased product (i.e. track access) was being delivered to an acceptable standard. The Office of the Rail Regulator also requires that a certain quality and quantity of rail infrastructure is provided by Network Rail and agrees with them the funding that the infrastructure owner will receive to deliver this (ECMT, 2005). The ORR has the power to fine Network Rail if it fails to meet the required standard.

### 4.2.2 The ‘Schedule 8’ Performance Regime

Internally the performance of the rail network is subject to a compensation regime called Schedule 8. This details the compensation that Network Rail has to pay an operator for ‘poor performance’ (i.e. below a set bench-mark) and the reward it receives for ‘good’ performance (i.e. above a set bench-mark). Preston, J., Wall, G., Batley, R., Ibáñez, J. N., and Shires, J. (2009) notes that Schedule 8 encourages Network Rail to provide good quality reliable infrastructure rather than being tempted to try and reduce costs. Train operators are also required to compensate Network Rail for any poor performance on their part (e.g. due to a train failure). The idea is to incentivise operators to provide reliable and properly resourced train fleets.

The ORR (2015, p3) notes that the key principles of the Schedule 8 regimes detailed in each Track Access Contract is to:

> (a) provide proper incentives to both parties to improve performance.
>
> (b) reasonably compensate operators for expected revenue loss and costs.
>
> (c) balance as far as possible risk and reward.
>
> (d) avoid perverse incentives and, in particular, ensure that through the performance regime Network Rail is not encouraged to discriminate unduly between users of the network; and
>
> (e) avoid undue constraints on the network or acting as a barrier to new entrants.

Schedule 8 operates as a ‘single-till’ mechanism. Schedule 8 minutes are divided into those that the Train Operator is deemed responsible for and those attributed to everyone else. The ORR (2015, p5) notes that:

> “any payment liability as a result of the impact of one train operator’s performance on another is channelled through what is called the ‘star model’ with Network Rail at its centre. The train operating company (TOC) payment rate is calculated so that, at the level of national performance across all service groups during the calibration period, Network Rail could expect to be compensated in full by the responsible TOCs for the payments it makes to the affected TOCs”.

In other words Network Rail pays compensation to affected train operators for all delays caused to their services other than those that they are responsible for. The calculated payment benchmarks being calculated on this basis.

However, delays to other TOC's trains caused by one TOC are charged for via Section 8 payments, and these will be higher on congested sections, whilst the capacity charges deter TOCs from proposing to run additional trains on congested sections, where they will not only incur Section 8 payments but receive such payments when delayed by other TOCs. Section 8 payments do not therefore, in themselves, provide sufficient disincentive to TOCs adding to congestion on already congested links.

Schedule 8 is a liquidated sums regime meaning that compensation rates are in accordance with a pre-determined fixed formula. The benchmarks are based on a set number of minutes lateness. Cancelled services are accounted for by assuming an equivalent number of minutes lateness. The benchmark is intended to represent an acceptable level of service and is arrived at following agreement with the ORR. The ORR notes that the Schedule 8 regime is intended to be financially neutral when all parties are performing in line with expectations (ORR 2012b).

Compensation rates vary between operators and flows and reflect the fact that delays to a heavily used peak commuter train causes greater cost than a lightly used rural service.

The Schedule 8 regime provides a real quantifiable cost (albeit internal to the rail industry) for the quality of performance of the rail network in Britain. Analysis of the financial statements from  Network Rail   for Control Period 4 (2009/10 to 2013/14) show that payments under Schedule 8 significantly exceeded the income . Only the first year of the period resulted in  a net (albeit modest) income[12]. This demonstrates that performance delivery by Network Rail, in terms of the total level of delays it is held accountable for, is currently falling short of the expected standard.

### 4.2.3 The Public Performance Measure (PPM)

The PPM has already been referred to in Chapter Three. As described earlier it is based on levels of 'Punctuality' and 'Reliability'. Its purpose is to provide information on the delivery of the timetable to its end customers (passengers and companies that transport their goods by rail). It was introduced in the late 1990s in Britain as a means of expressing the quality of service delivered to the public. Originally used to monitor the performance of passenger operators, a similar measure has since been introduced for freight operators.

[12] See Table 4.4

Figure 4.1 shows the overall PPM for franchised passenger operators between 1997/98 and 2012/13. Average Annual PPM was 89.8% at the start of the period, but declined sharply following the Hatfield rail crash in the year 2000 due to the widespread imposition of temporary speed restrictions. Performance then gradually improved to reach 89.9% in 2007/8. It has then fluctuated between 90.5% and 91.5% reaching a high of 91.6% in 2011/12. The period ended with a PPM of 90.9%.

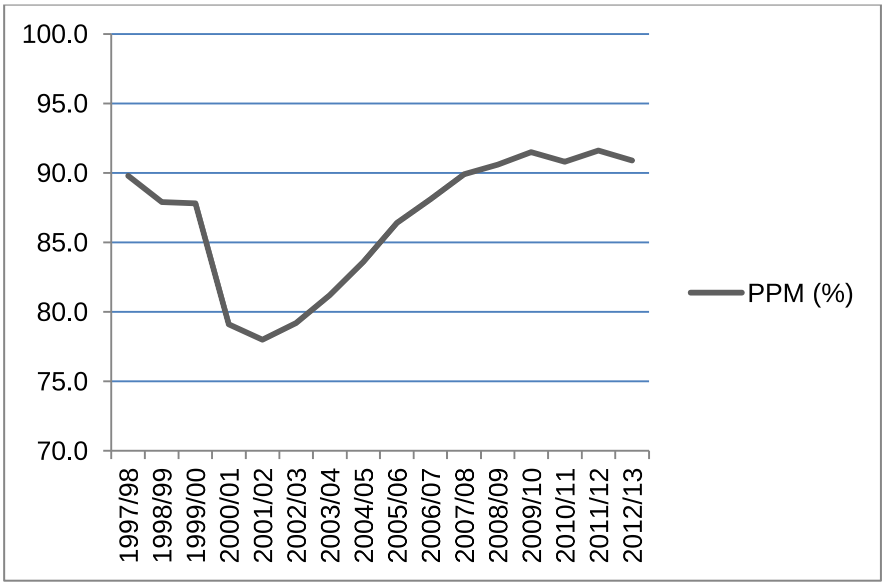


**Figure 4.1** PPM for franchised passenger operators - 1997/8 to 2012/13 (ORR National Rail Trends Portal, 2013a).

Nichols consultants in their review of Network Rail’s performance plans for Control Period 5 note that the Secretary of State for Transport has specified that PPM in England and Wales, should achieve an overall level of at least a 92.5% moving annual average by the end of the period. Interestingly, the minister “wishes to have a higher level if the ORR determines this is value for money and can be affordably achieved without compromising delivery of other ...requirements”. (Nichols Group, 2013, p11).

Network Rail, themselves, highlight that performance is a competing priority with capacity and cost observing that “as demand continues to grow on key parts of the network, in some places it is no longer possible to simultaneously cut costs, increase capacity and deliver more trains on time. Often one comes at the price of another” (Network Rail, 2013a, p5).

### 4.2.4 Two Examples of Performance Incidents

Before exploring the industry's perspective on the reasons for current levels of railway performance, it is useful to consider two example performance incidents provided by Network Rail (2013d).

A signal failure on 15th June 2011 at Watford Junction caused 6,763 minutes of delay. Of this 3,687 minutes (or 55%) were classed as Primary Delay; 2,150 minutes (or 32%) were classed as Congestion Related Reactionary Delay and the remaining 926 minutes (or 13%) were classed as Late Start Reactionary Delay.

A track circuit failure on 14th June 2012 near Ashchurch (between Cheltenham and Worcester) caused 4,390 minutes of delay. Of this 951 minutes (or 22%) were classed as Primary Delay; 2,902 minutes (or 66%) were classed as Congestion Related Reactionary Delay and the remaining 537 minutes (or 12%) were classed as Late Start Reactionary Delay

It can be seen that the two different incidents have different proportions of delay type. This is unsurprising and will be a function of the location, duration and type of primary incident. Earlier analysis by Preston, J., Wall, G., Batley, R., Ibáñez, J. N., and Shires, J. (2009) suggests that overall reactionary delay accounts for approximately 60% of total delay minutes with CRRD making up 40% of the total.

One startling aspect of both the incidents is the spread of the resulting reactionary delay across the British rail network. For example, Network Rail's analysis of the Ashchurch incident shows CRRD as far away as the route between Edinburgh and Glasgow, hundreds of miles away from the performance incident in the west of England.  This propagation of delay underlines the inter-connected nature of Britain's rail network.

### 4.2.5 The Rail Industry's View

It is also useful to consider the rail industry's views on the general reasons for 'poor' timetable performance.

In 2011 the ORR issued an enforcement order to Network Rail requiring it to deliver a plan to improve its performance of the long-distance train sector in the 2012/13 period.  ORR's (2012c) review of the plan Network Rail produced gives an illuminating insight into the Regulator's view of some of the factors leading to poor performance. These included:-

- Traffic Growth (above the amount planned for by Network Rail).
- Delay per incident (which has risen).

- External factors (with a rise in delay due to fatalities and trespass being specifically mentioned).
- Severe Weather in 2009/10 and 2010/11.
- Track Quality (with an increase in unplanned temporary speed restrictions).
- The restructuring of maintenance delivery by Network Rail.
- Timetabling including problems with delivery of a new timetabling system.
- Delays in key projects.
- Train Operator PPM failures.

More recently, Network Rail in conjunction with consultants undertook a study (Network Rail 2013g) aimed at improving their understanding of the relationship between performance ‘inputs’ and ‘outputs’ (particularly PPM). This looked in detail at the experience of services provided by two train operators (South Eastern and East Coast Trains).The conclusions reached reinforce and add to the list of factors already given. They can be summarised as follows:-

- Increasing the number of trains will generally worsen performance.
- However, the impact of additional trains can be negated by timetable improvements where they reduce complexity.
- Timetable complexity is a significant factor in performance (particularly the mix of ‘fast’ and ‘slow’ trains and the number of crossing moves at junctions).
- There is generally worse performance on long-distance routes. It is suggested that this in part is due to the greater distances travelled and the consequent greater risk of incurring delay and the increased interaction with other services.
- Performance can differ markedly by direction (this is explained as an increased complexity towards termini as services converge and the fact that the absolute level of delay will be greater towards the end of a journey).
- Increasing termini capacity utilisation will worsen performance.
- The quality of the timetable is a key driver of reactionary delay (a definition of timetable quality is however not given).

It is clear that the lists reinforce the findings of the literature review discussed in Chapter three. In particular, the volume of trains is listed by both the ORR and Network Rail as a key factor in the level of performance. However, whilst many of the ORR's factors refer to the reasons behind the original performance incidents; Network Rail's list concentrates on the nature of capacity utilisation itself. From the point of view of this thesis Network Rail's list is therefore the more useful of the two. It can be seen for example that both timetable complexity and distance travelled appear on the list.

## 4.3 Theory of Congestion Charging

Before describing how the Capacity Charge was introduced and subsequently developed in the UK, it is useful to first outline the basic theory of congestion charging.

Rouwendal and Verhoef (2006, p107) in their paper on the basic economic principles of road pricing, refer to Pigou and Knight's work in the 1920s as the foundations for the concept of congestion charging. The latter referred to a greater increase in Marginal Cost compared with Average Cost that resulted from the addition of an additional vehicle to a congested road. The Average Cost (or AC) experienced by an individual driver will increase as the road becomes more crowded due to a consequent reduction in the average speed. However, this cost only accounts for his personal increase in journey time. Only by summing together all the individual increases in journey time can the full impact of congestion, or the marginal cost (the MC), be understood. With each additional trip, there are more vehicles to be affected by any delays but also to affect each other and therefore as congestion increases the difference between the MC and the AC rises sharply. This relationship is illustrated in Figure 4.2.

It can be seen that the shape of the curve for the increase in Marginal Cost as traffic flow increases has a striking similarity to the relationship between rail capacity utilisation and reactionary delay described in Chapter 3 (as shown for example in Figure 3.10). Indeed, the exponential relationship observed by many researchers is clearly due to the increase in Marginal Cost as rail congestion increases.

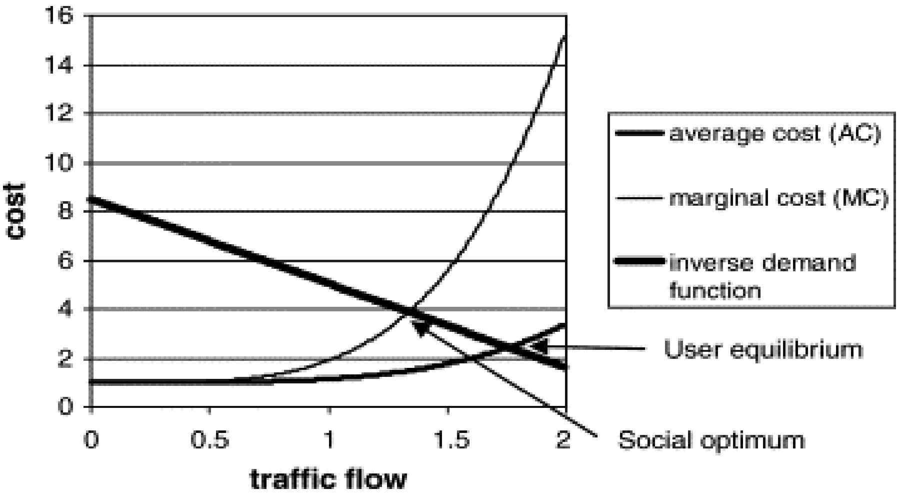


**Figure 4.2** The Pigovian-Knight relationship (source: Rouwendal and Verhoef, 2006, p107).

The Pigou-Knight solution was to propose a corrective tax based on the difference between the marginal cost and the average cost of additional traffic. In doing so, each vehicle is then faced with their share of the full cost of congestion (i.e. their share of the corrective tax plus their personal increase in average cost). By introducing the tax it was then theorised that traffic demand would reduce from a user equilibrium (where the demand function crosses the AC line) to a social equilibrium (where the demand function crosses the MC line).

“The prescription that prices should equal marginal costs is probably among the best known policy advices of economists” (Rouwendal and Verhoef, 2006, p108). This included Mohring (1970) who stated that making price equal to short run marginal costs is one of the necessary prerequisites for the efficient utilisation of any given level of fixed capital plant. This approach is referred to as ‘first best’ pricing which in a transport network means that at least all other congested routes are tolled as well, otherwise distortions in the incentive effect begin to occur (Rouwendal and Verhoef, 2006). In some cases, it makes sense to lower the toll below the optimum marginal cost level to prevent too much traffic diverting onto untolled and parallel routes and causing too much congestion. This is one example of ‘second best’ pricing.

## 4.4 The Capacity Charge in Britain

### 4.4.1 Development of The Original Charge

Britain’s Capacity Charge is explicitly linked to the concept discussed in chapter three that increased capacity utilisation will lead to increased reactionary delays.

Prior to 2002, Railtrack negotiated on a case-by-case basis increases to fixed access charges to recover the additional congestion costs expected to arise following the introduction of new services. However, this approach was not considered to be transparent or predictable enough and the development of a tariff based congestion charge was intended to address this (Thomas and McMahon, 2005). The original charge also only applied to new access rights. It was felt that this approach “... provides no signal to existing users about congestion costs” (Symonds Group Ltd, 2000, p6). “With much of the ....network at or near capacity, this absence of signals encouraging the efficient use of the network was a significant cause of concern” (Gibson, S., Cooper, G. and Ball, B.  2002, p342).

The tariff based Capacity Charge was therefore introduced in 2002, following the Access Charge Review in 2000. Gibson, S., Cooper, G. and Ball, B. (2002) note that the aim of the Charge was to both help the Infrastructure Manager recover the expected increase in marginal congestion costs arising from accepting more traffic onto the network (so that they received an incentive to do so) and to send out appropriate price signals to stakeholders which would encourage the efficient utilisation of available capacity. However, Faber Maunsell (2007) as part of the CP4 review of the Capacity Charge felt that there was a potential conflict between these two objectives in that one required the Charge to have a high degree of granularity so that it was as cost-reflective as possible but the other required the charge to be sufficiently simple for it to be manageable and easily understood.

In order to produce the Capacity Charge, the British rail network was sub-divided into a large number of  Constant Traffic Sections. Gibson, S., Cooper, G. and Ball, B. (2002) note that the choice of subsequent geographic sections for the tariffs was aimed at retaining a high level of granularity in this dimension. This was in order to appropriately signal costs to operators to influence their timetabling decisions and implied including all important stations and junctions as tariff end points together with any junction where more than 25% of traffic ‘turned off’. The  model had approximately 2,750 geographic tariff cells and distinguished between direction of travel (Gibson, S., Cooper, G. and Ball, B., 2002, p349)

Temporal differences were assumed to be solely due to changes in capacity utilisation on each route section. A total of 13 timebands were used in the original analysis and these are shown in Table 4.1. To reduce complexity the timebands remained constant across the network.

**Table 4.1** The Original Capacity Charge Timebands (Source: Gibson, S., Cooper, G, and Ball, B., 2002, p349).

| Weekday | Saturday | Sunday |
|---|---|---|
| 00:00 – 05:00 | 00:00 – 05:00 | 00:00 – 09:00 |
| 05:00 – 06:30 | 05:00 – 08:00 | 09:00 - 24:00 |
| 06:30 – 09:30 | 08:00 – 18:00 | |
| 09:30 – 16:30 | 18:00 - 24:00 | |
| 16:30 – 19:30 | | |
| 19:30 – 21:00 | | |
| 21:00 – 24:00 | | |

Regression analysis was then used to establish a significant relationship between capacity utilisation and timetable performance. A number of functional forms were tested, but once again an exponential curve was found to best describe the relationship. Equation (18) shows this specification (Gibson, S, Cooper, G. and Ball, B. , 2002, p347):-

$$D_{it} = A_i * \exp(\beta C_{it}). \qquad (18)$$

where

$D_{it}$ is the CRRD per train mile on geographic section i in time period t;

A is a section specific constant;

β is the coefficient of the capacity utilisation which varies by route and

C is the capacity utilisation calculated for section i and time period t.

It can be seen that CRRD per train mile[13] was used as the dependent variable ($D_{it}$) in the specification whilst link-only CUI was used to calculate the explanatory variable ($C_{it}$). The purpose of the section specific constant (*A*) was intended to capture any spatial differences between individual sections (e.g. the influence of an adjacent major node on reactionary delay). In a similar fashion the coefficient β was intended to capture any differences at a route level. Twenty-four strategic rail routes were used (e.g. ECML, WCML, Northern Transpennine).

The tariff for each cell was then calculated on the basis of the additional reactionary delay 'produced' by one extra average[14] train. The equation is shown below as Equation (19) (Faber Maunsell, 2007, p10): .

$$\Delta\, D_{it} = A_i * \exp(\beta^* C_{it}') - A_i * \exp(\beta^* C_{it}). \qquad (19)$$

Where

$\Delta\, D_{it}$ is the increase in CRRD per train mile for geographic section i and time band t.

C' is the new CUI value following the addition of one average train.

The process is illustrated in Figure 4.3.

This calculated increase in CRRD per train mile for each cell was then used to calculate individual  tariffs. Faber Maunsell (2007, p10) in their review of the Capacity Charge for CP4 explain that this was based on the average cost of a minute lateness for that geographic section using Schedule 8 rates (which had to be weighted due to the different rates for different operators and the fact that they apply at Service Code level). The tariff was also reduced to the proportion of delay that the infrastructure manager was historically responsible for (the 'Fault Percentage') for that geographic section. In other words the tariff excludes the delay that a train operator causes to its own services (their Average Cost) but includes the cost to everyone else (the Marginal Cost minus the AC). The Capacity Charge is therefore based on the Pigouvian-Knight approach.

[13] CRRD per train mile represents a more than linearly increase in reactionary delay and thus shows any increased marginal cost due to congestion ( Faber Maunsell 2007, p9).

[14] In fact this is the capacity utilisation for a particular cell increased by n+1 / n (Arup, 2013, p27).

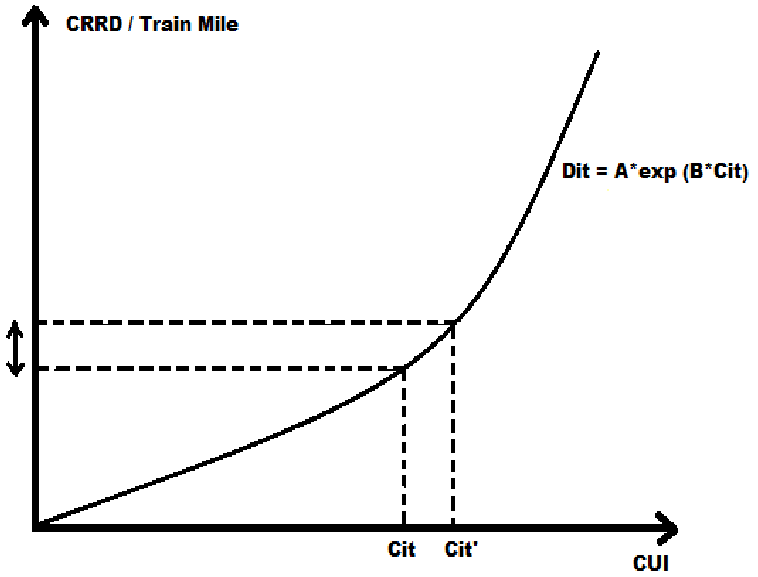


**Figure 4.3** The Expected Increase in CRRD per Train Mile Following an Increase in CUI (Source: Faber Maunsell, 2007, p10).

Multiplying the tariff by the number of trains for that section and time-band then gave a total corrective tax for each cell.

A number of comments can be made about these adjustments to the Capacity Charge. Firstly, the multiplication of the calculated increase in CRRD by the average Schedule 8 rates provides the necessary monetary element to the Charge. Obviously, Schedule 8 rates therefore have a significant effect on the size of the charge and the overall income. Indeed it is noted later in this thesis (p87) that in the case of Franchised Passenger Operators this factor was expected to represent almost the entirety of the increase in the Charge in CP5. It is outside the scope of this thesis to discuss the calculation of the Schedule 8 rates themselves. However, it is useful to discuss the use of a weighted average Schedule 8 rate. Although, delays will be suffered by different trains on a section this is impossible to predict in advance. The use of a weighted average rate therefore can be considered the most sensible option.

The application of a Delay:Lateness ratio is a necessary reflection of the fact that the calculation of the increase in the impact of congestion is in terms of delays and  the basis of Schedule 8 is lateness.

The application of the Fault Ratio also makes sense but this time from an economic point of view. As noted it replicates the Pigouvian-Knight approach. Additionally excluding TOC on self delay, although meaning that not all potential delay is charged for, means that Network Rail does not double charge for the delays that Train Operators cause to themselves. For these reasons the Fault ratio as it has been applied appears sensible.

One simplification of the charge was to introduce a ‘de-minimis’ threshold where tariffs below 10p per mile were reset to zero. Gibson, S., Cooper, G. and Ball, B. (2002, p351) note that this significantly reduced the complexity of the charge as 77% of the individual tariffs were subsequently reduced to zero. The economic principle of not charging for certain routes however needs to be considered. As will be discussed in greater detail later in the thesis, the potential consequence of de-minimis is to increase the incentive to switch to less congested routes above that implied by a true equilibrium. The risk is that the consequent transfer of traffic is above a level which is sensible, from a capacity point of view on the alternate routes.

Capacity charge rates were also halved by the ORR at a late stage in the review as it was felt that if operators faced the full marginal costs of the services they operated then “higher access charges would reduce the growth of rail services on the network and this would conflict with government growth targets” (Gibson, S., Cooper, G., and Ball, B., 2002, p351). It was determined that the infrastructure manager would recover the other half of expected congestion costs from the Strategic Rail Authority[15].

The halving of the Capacity Charge for this reason shows a conflict between a desire to promote growth at a certain level and the objective of sending appropriate signals for the efficient use of capacity on the network. Since the latter can said to have an objective of applying restrictions on growth. The two objectives are potentially mutually exclusive. It can be argued that the halving of the charge helped ensure that the ‘rationing’ of access to congested routes could not be effective. It can also be argued that in taking this action the ORR undermined a key objective of this charge. Perhaps a better approach would have been to take a selective view where a full 100% Capacity Charge was appropriate and where it might ‘be relaxed’ in order to meet the objective of encouraging growth.

A sample Capacity Charge tariff using the original approach is conveniently provided by Gibson, S., Cooper, G. and Ball, B. (2002) for three different time bands for the Midland Main Line between Sheffield and London St. Pancras. A part of this is reproduced as Table 4.2. It can be seen that there

[15] A public body in existence between 2001 and 2006.

is a clear price differential between the three time-bands and clear difference in the tariffs between the individual geographic sections. The highest tariffs can be seen between Sheffield and Chesterfield in the AM peak, whilst between Clay Cross South Junction and Mansfield Junction many of the cells have a zero charge. Interestingly, the off-peak band is not always the cheapest charge with Mansfield Junction to Nottingham in particular having its highest tariff in this timeband.

**Table 4.2** Extract of Original Capacity Charge Tariffs (£ per train mile) (from Gibson, S., Cooper, G. and Ball, B., 2002, p353)

| From | To | AM Peak | Off peak | PM Peak |
|---|---|---|---|---|
| Sheffield | Dore Station Jn | 1.95 | 0.75 | 0.65 |
| Dore Station Jn | Chesterfield | 1.15 | 0.35 | 0.45 |
| Chesterfield | Clay Cross Sth Jn | 0.35 | 0.15 | 0.25 |
| Clay Cross Sth Jn | Trowell Jn | 0.00 | 0.00 | 0.00 |
| Trowell Jn | Radford Jn | 0.25 | 0.00 | 0.00 |
| Radford Jn | Mansfield Jn | 0.25 | 0.00 | 0.00 |
| Mansfield Jn | Nottingham | 0.35 | 0.65 | 0.55 |

As discussed the Capacity Charge in its original form was therefore based on the marginal cost of one additional average train and disaggregated by location and time of day.

Gibson, S., Cooper, G, and Ball, B. (2002, p353) noted their belief that the Capacity Charge would provide an “appropriate incentive to (the infrastructure manager) to make efficient decisions over use of the network in its timetabling decisions and compensate it for the marginal congestion costs incurred ... the capacity charge will ... provide a signal to operators... over and where the network is congested, and should therefore influence operator decisions towards efficient requests for track access in timetable bids and signal where investment is required”.

### 4.4.2 Implementation of the Capacity Charge

Faber Maunsell (2007) in their review of the Capacity Charge for Control Period 4, revealed that in-fact the charge had been implemented for franchised passenger operators by their Service Groups. There are approximately 130 Service Groups in total (Network Rail 2012c, p14) so this represents a very considerable aggregation of the original tariffs. This was due to “a number of implementation issues related to billing”‘ (Faber Maunsell, 2007, p3). Therefore, rather than tariffs which differed by time-band and location as had been the original intention; single tariffs for the entire journey of every train in the limited number of Service Groups were applied. These had been produced for the weekday rates by weighting all the applicable time and location tariffs that had been calculated. A distinction was made between weekday and weekend operation by then applying a flat 25% discount to the latter.

**Table 4.3** Example Capacity Charge Rates (CP4 2009/10 Prices) (Source : Network Rail 2008b).

| Franchised Passenger Train Operator | Service Group | Weekday rate (£/train mile) | Weekend rate (£/train mile) |
|---|---|---|---|
| National Express East Coast | HB01 | 0.4143 | 0.3107 |
| National Express East Coast | HB02 | 0.4980 | 0.3735 |
| National Express East Coast | HB04 | 0.4143 | 0.3107 |
| National Express East Coast | HB05 | 0.4143 | 0.3107 |
| National Express East Coast | HB99 | 0.1838 | 0.1378 |

An example of the charges introduced is shown in Table 4.3. It can be seen that although there are five Service Groups, in real terms there are only three weekday and three weekend tariffs which apply to all National Express East Coast trains including the movement of empty coaching stock. As an example, all services between Leeds and London, which are included in Service Group HB02, were charged a flat weekday rate of £0.498 per mile irrespective of the time of day they operated, how congested a particular location on the route was calculated to be or the direction of travel.

The application of a flat weekend discount to the weekday rates which was applied to the original and recalibrated Capacity Charge also needs to be questioned. The assumption is that the change between weekdays and weekends is uniform across the network. This will clearly not be the case. For example, main lines such as the ECML with heavy weekday commuter traffic will have a different character to rural lines where the traffic is more likely to be associated with social welfare. The assumption that the weekday peaks will also exactly apply to the weekends will also not be the case. The impact is therefore expected to be that the weekend charge is too imprecise and therefore unlikely to achieve the stated objectives. It seems unclear other than for convenience why after calculating weekend rates an adjustment was instead made to the weekday rates. For the reasons discussed above it is believed that a much better approach is to produce separate tariffs for the weekends. Freight services were subject to a single 'flat' tariff no matter the time of day and routing. Once again a single weekend discount of 25% was applied. However, to reflect the greater flexibility associated with pathing these services a 10% discount was also applied to the tariffs. ( Network Rail, 2012c, p16).

The rationale behind the application of a Freight Flexibility discount is that since paths are easier to time into less congested periods they should receive some acknowledgement of this benefit. However, once again this flat rate application of a discount appears rather crude. It does not distinguish between freight services that in fact have fixed rights and are consequently much less easier to retime than those with contingent rights. The universal application of this flexibility discount means that there is no incentive for freight operators to be agreed to be retimed. If such a discount were to be an effective incentive there would need to be some mechanism whereby the discount was applied following agreement by an Operator to retime to a less congested route or time. The practicality of this is debatable and therefore the application of different tariffs for time and geography remains the preferred option.

Faber Maunsell's (2007) review for CP4 used the same methodology for calculating the Capacity Charge. The intention was to update the relationships with the latest traffic and CRRD data. The number of time-bands was however reduced from the original 13 to 6 (Weekday Off-Peak, Weekday AM peak. Weekday Inter-Peak, Weekday PM Peak, Saturday and Sunday). 6,000 Constant Traffic  Sections were used (Faber Maunsell, 2007, p7) which were aggregated into 600 geographic tariff sections.  The intention with the CP4 review was to move to this level of disaggregation from the Service Group level. Faber Maunsell comment that “Network Rail’s billing processes have developed since the Capacity Charge was first introduced, and it is reasonably certain that this level of granularity can be implemented” (Faber Maunsell, 2007, p17).

However, Network Rail in their consultation for the CP5 review of the Capacity Charge made it clear that in fact the Charge had continued to be levied on the basis of Service Groups with ‘billing issues’ again being given as a reason (Network Rail, 2012c). In addition, the new relationships developed for CP4 were not used. Instead, the CP4 tariffs were based on the original 98/99 timetable and performance data with performance payment rates from 2004/5 which were then updated on an annual basis to take into account increases in RPI (Arup, 2013, p41).

### 4.4.3 How Successful has the Capacity Charge been in Meeting its Objectives?

To recap, the two main objectives of the Capacity Charge were to:-

- recover the additional marginal cost to the infrastructure manager of accepting more traffic onto an already crowded network. This would mean it wasn’t dis-incentivised from accommodating traffic growth.
- send out price signals to the infrastructure manager, train operators and other stake-holders to encourage more efficient use of the existing rail network.

Evidence from the application of the Charge, and in particular the comments made as part of Network Rail’s consultation process for the CP5 re-calibration exercise, provide a useful starting point in considering how successful the Charge has been in meeting these two objectives.

An interesting comparison is to examine the difference between Schedule 8 performance payments and Capacity Charge income. Table 4.4 compares Network Rail’s Capacity Charge income with its Schedule 8 payments for the five years of Control Period 4 for franchised passenger services.  Network Rail note that 97% of the Capacity Charge is paid by franchised passenger services (2012c, p5) so the payments potentially provide a very good indication of  whether the Capacity Charge successfully recovered the increase in marginal cost due to accepting more traffic onto the network.

Table 4.4 shows that on a national level, with the exception of the final year of CP4, income from the Capacity Charge considerably exceeded payments under the Schedule 8 regime. Since the Schedule 8 payments include both primary and reactionary delays for which Network Rail is responsible for, the discrepancy is even greater than implied by Table 4.4. Over the five years of CP4, Capacity Charge income was nationally 81% higher than Schedule 8 payments.

**Table 4.4** Comparison of Capacity Charge and Schedule 8 Performance Regime Payments for CP4 (Prices in £m for year in question) (Compiled by author from: Network Rail, 2014c; 2013b; 2012b; 2011a and 2010b).

| **Charge** | **2009/10 (£m/yr)** | **2010/11 (£m/yr)** | **2011/12 (£m/yr)** | **2012/13 (£m/yr)** | **2013/14 (£m/yr)** |
|---|---|---|---|---|---|
| Total Fixed | 782 | 912 | 887 | 1,109 | 1,464 |
| Variable Usage | 137 | 137 | 150 | 160 | 166 |
| Traction Electricity Charge | 227 | 218 | 200 | 236 | 267 |
| Electricity Asset Usage Charge | 8 | 8 | 9 | 10 | 10 |
| **Capacity Charge** | **156** | **158** | **169** | **177** | **183** |
| Schedule 4 Net Income* | 188 | 167 | 178 | 149 | 146 |
| Schedule 8 Net Income* | 3 | 3 | 0 | 0 | 0 |
| Total Franchised Income | 1501 | 1603 | 1593 | 1841 | 2236 |
| **Schedule 8 Payments** | **2** | **(56)** | **(80)** | **(136)** | **(197)** |

(* Passenger Charge Access Charge Supplement).

However, this assumes that it is appropriate to make a direct comparison between Schedule 8 and Capacity Charge payments. Schedule 8 and the Capacity Charge are clearly complimentary, especially given the common basis of the actual monetary value. However, as noted previously (p63) the Schedule 8 regime is intended to incentivise and improve current performance. In contrast, as previously discussed, the Capacity Charge is designed to compensate the infrastructure manager for the expected performance impact due new traffic increasing congestion. At the same time the Capacity Charge is intended to incentivise all parties to improve the efficiency with which capacity is used on the network. Although linked, the two regimes clearly therefore have different functions.

Furthermore, as noted in point (e) on page 62 of this thesis the Schedule 8 regime is not intended to provide undue constraints on the network or to act as a barrier to new entrants. Whilst train operators who cause delay make payments through the regime, those who suffer delay receive compensation. Providing the performance targets are met the regime is intended to be financially neutral. Although, penalties are likely to be higher on congested parts of the network so will the level of compensation. The Schedule 8 regime is therefore not designed to provide sufficient incentive to operators to use less congested parts of the network. In contrast by seeking to incentivise a more efficient use of capacity, it could be argued that the Capacity Charge is intended to act as a form of constraint due to its objective of limiting the impact of congestion. To achieve this objective the Capacity Charge needs to be set at a sufficiently high level. Finally, all Capacity Charge payments are to Network Rail in contrast to the two-way nature of the Schedule 8 ‘star model’.

The overall Schedule 8 cost to Network Rail in Table 4.4 will not therefore directly equate to the overall Capacity Charge income shown. In summary, therefore Schedule 8 payments which are designed to compensate and incentivise current performance cannot be directly compared with Capacity Charge payments which are designed to compensate and incentivise the future use of congested routes.

Furthermore, the ORR (2013b) notes that Network Rail’s results show an overall under-recovery of costs from freight operators through the Variable Usage Charge. There is therefore an argument that the overall effectiveness of the Capacity Charge should not be judged in this way.

For these reasons, although an interesting comparison, the evidence presented in Table 4.4 does not answer the question about the effectiveness of the Capacity Charge. However, a number of comments have been made about the substantial size of the charge and why it may be over-recovering the marginal cost on a national basis.

A number of possible factors have been suggested for this large discrepancy:-

- Firstly, the belief that over-recovery was due to the application of the tariffs to all traffic rather than just incremental trains (for example DB Schenker, 2012 and G.B.Rail Freight, 2012). This contrasts with the principle behind the Capacity Charge, discussed earlier, that all trains should be exposed to the cost of congestion in the belief that all operators are then subject to the "economically correct price incentives and signals" (Network Rail 2012c, p11).Network Rail themselves believe that just applying the charge to incremental trains could give "economic advantages to incumbent operators and services. This would be contrary to relevant legislation and could stifle competition in the rail market" (Network Rail, 2013c, p36).
- Secondly, the application of a flat-rate tariff to Service Groups suggests that congested parts of the network are being under-charged and non-congested parts of the network are being over-charged. If the latter outweighs the former in terms of total train miles then the overall result will be an over-charging.
- Thirdly, Centro (2012) make the point that the charge implies an increase in traffic always means an increase in congestion costs. They suggest that if the increase is coupled with more efficient capacity utilisation congestion costs may reduce.
- Finally, Arup (2013) refer to a declining trend in the level of CRRD per train mile. Therefore, over time there has been a change in the relationship between capacity utilisation and marginal cost. They also note that there has been a reduction in primary delays. The CP4 tariffs are therefore based on 'worse' performance assumptions than actually occurred.

These comments therefore do suggest that there is an issue with the size of the charge and how it has been calculated. The question therefore is not how the Capacity Charge performs compared with the Schedule 8 regime, since as noted previously they fulfil different roles but instead whether the Capacity Charge is appropriately calculated for the function it is intended to perform.

Interestingly in the case of franchised passenger operators, Network Rail does not benefit directly from the discrepancy in charges. This is because for the life of their franchises, operators are protected from variations in charges through an adjustment to their Fixed Access Charge (Network Rail, 2012c). Only freight companies, Open Access passenger operators and funding bodies are subject to the commercial risk of being over-charged. The Rail Freight Group (2012, p2) estimated that over-recovery via the Capacity Charge in CP4 up to 2012 was £12 million and therefore a “significant issue”.

Network Rail (2012c, p5) themselves refer to the concern that “the charge does not always fully compensate [them] for the increased performance risk associated with accommodating new services”. Therefore, whilst the Charge may over-recover the marginal cost on a national basis as evidenced by the significant level of income shown in Table 4.4, it clearly does not always recover the local marginal cost associated with accommodating specific new traffic. This reinforces the view that the Capacity Charge may over-charge on some parts of the network but under-charge on others. Additionally, Freightliner (2012, p8) noted the Capacity Charge’s objective in preventing NR being dis-incentivised from accommodating additional traffic but their experience was that “local NR staff are reluctant to agree to new services as they are seen as a perceived risk to their performance targets”.

In terms of the objective of encouraging more efficient use of the network through effective price signals, one of the key issues with the Capacity Charge is the use of a ‘flat-rate’ tariff for operators. Theoretically a congestion charge differentiated by time and location provides the most effective incentive. However, AECOM consultants (2012), despite noting that at least a division between peak and off-peak services would be a good method of incentivising efficient use of the network, suggested that anything but a ‘flat’ rate tariff might actually produce a perverse incentive. This was on the basis that operators might be encouraged to ‘cluster’ services at the margins of cheaper tariff bands. Network Rail (2013c) themselves also referred to the possibility of band ‘clustering’.

Network Rail (2013c) also expressed concern that there was a risk during the timetable development phase that services could be ‘flexed’ into higher rate time bands. G.B.Rail Freight (2012) in their response believed that geographical differentiation could lead to a perverse incentive of freight traffic being encouraged to use ‘unsuitable’ routes.

Another important point which emerged during the consultation process for the CP5 recalibration was raised by freight companies or freight stakeholders. There was a strong desire to keep the Capacity Charge as low as possible for Freight traffic since it is “seen as a surcharge by [their] customers” (Freightliner, 2012, p2) and “most rail freight sectors are highly elastic so that increases in the level of charge could lead to traffic reversion to road” (Rail Freight Group, 2012, p1). This should be avoided as rail freight saves £722 million per annum in road congestion costs (Freightliner, 2012, p2) with it being suggested that the proposals were not aligned to ORR’s duty “to promote carriage of goods by rail” (Freightliner, 2012, p2). This raises the possibility that Freight traffic at least should be subject to ‘second-best’ pricing. Indeed, De Palma and Lindsey (2011, p1382) note that pricing discounts are sometimes offered to groups for “public acceptability reasons” with the 90% discount offered to residents within the London Congestion Charging Cordon being given as one example.

However, there was also a great deal of support in the CP5 consultation responses for tariffs that did vary by time of day and location. PTEG (2012), who support the six Passenger Transport Executives in England, believed it was inefficient to levy uniform charges across the day and also recommended disaggregation by route section. Transport for London (2012) did not agree to a single tariff for freight as this did not take into account congestion. They noted that “freight services operate on the congested North London Line and should pay a higher tariff for routes such as this than they do on uncongested routes” (Transport for London, 2012, p2). Network Rail however continued to support a ‘flat’ rate for freight as they believed it did not lead to “undue discrimination” was “practicable” and provides “certainty” (Network Rail, 2013c, p16). Centro (2012, p3) were particularly concerned about the lack of time and geographic differentiation of tariff rates. They considered it “wrong” and “economically inefficient” that they should be penalised for trying to fund services at quiet times when there is currently inadequate provision. They also noted instances of off-peak services having to be withdrawn due to the cost of the flat-rate Capacity Charge.

It is clear that the Charge has not worked entirely in the way it was intended, has not always met its objectives and has produced some perverse incentives. It is also unclear how it is able to send out effective price signals without some form of time and geographic differential between tariffs. However, the issue of appropriate granularity versus complexity of implementation is one of the key messages that emerges. In seeking to achieve a balance between the two, it can be said that the Capacity Charge is not optimised on economic principles.

### 4.4.4 Recalibration of the Capacity Charge for CP5

For the Control Period 5 recalibration of the Capacity Charge essentially the same methodology was adopted. The opportunity was however taken to update the capacity utilisation and performance data and the Schedule 8 payment rates. The latest regression techniques were also used. The work provides an excellent framework for the analysis undertaken for this thesis. The methodology used is therefore described in greater detail in the next chapter.

Link-based CUI was retained as the measure of capacity utilisation. Network Rail (2012c) as part of the initial consultation had referred to 'statistical noise' in the previous analysis and expressed the view that other determinants of reactionary delay (and in particular junction and station capacity utilisation) could be considered. However, in the end the same approach was adopted as before. This was for consistency and recognised that CUI was the accepted standard for measuring capacity utilisation in Britain (Network Rail, 2013c). Network Rail (2013c) did recognise the view, expressed by some consultation respondees, that CUI was not an ideal metric and suggested that this was something that could be revisited for future recalibrations. CRRD per train mile was retained as the measure of the cost of congestion. The approach to calculating the speed-flow relationship therefore remained the same as that used for the original work.

CUI and CRRD per train mile values were calculated for individual 'links' and time-bands. For this analysis a new set of time-bands were employed. These divided each day into three hour periods (rather than using time bands of irregular duration). For example, 0700 to 1000 hours, 1000 to 1300 hours and 1300 to 1600 hours. This gave a total of twenty-four time-bands (taking into account Weekdays, Saturdays and Sundays).

Imperial College London were contracted to assist with the econometric analysis. They undertook a thorough analysis of the relationship between CUI values and CRRD per train mile. The likely functional form was first of all established using semi-parametric modelling (Imperial College London, 2013). The results of this were then used to identify a number of functional forms for the subsequent regression analysis. As noted, this analysis was used as the framework for the analysis carried out for this thesis and is therefore described in detail in the next chapter.

The analysis established a relationship between capacity utilisation and performance. Once again the Exponential functional form[16] was identified as the most appropriate (Arup, 2013). However, although Arup were happy with the strength of the relationships, Imperial College London (2013, p31) did raise a number of concerns. They referred to three particular issues which potentially could lower the strength of the relationship:-

- Endogeneity bias due to network effects, i.e. the capacity utilisation in one section was affecting the capacity utilisation and reactionary delay in another part of the network.
- Endogeneity from reverse causality between CUI and CRRD, in other words the expected levels of CRRD on a specific link at a specific time of day were influencing timetable preparation and thus the level of CUI.
- Endogeneity bias from omitted variables.

Inclusion of the ‘other’ variables described in Chapter Three are intended to help address these potential issues.

One important difference from the original calibration described by Gibson, S., Cooper, G. and Ball, B.(2002) was the use of a single network wide capacity coefficient β, rather than the 24 different ones based on strategic route sections. Unfortunately, no explanation is given by Arup (2013) for this decision. However, this has potential implications for the results as it means that the slope parameter (i.e.‘β’) is the same for all parts of the network. The relationship between capacity utilisation and CRRD is therefore assumed to be the same across all parts of the country apart from differences in the section specific element of the specification (i.e. ‘*A*’). However, in the original calibration the ECML and Wales and the Borders for example which might be expected to have different characteristics had different β values. Arup (2013, p22) did investigate the use of the 268 strategic route sections used by Network Rail, in an attempt to account for network effects, but this was on the basis of averaging CUI across each section and was not pursued.

A further point is that although the preferred relationship between capacity utilisation and reactionary delay was an exponential one, the slope (as expressed by ‘β’) was not particularly convex (in contrast to the expectation

[16] Equation (18)

shown in Figure 3.10 for example). Furthermore, during the analysis a number of variants of the data set were tested, from the full data set to one for example that excluded CUI and CRRD values of 0 or a CUI of greater than 100[17]). This reduction significantly affected the value of β from 0.00062 to 0.00025. The decision was then taken to adopt the more 'conservative' value from the reduced data set[18]. Arup (2013, p29) note that "it should be recognised that the 'true' slope parameter is likely to be greater than the figure used in this analysis, and that it may be appropriate to review the estimate of beta in the future".

Once again the tariffs for each time-band and geographic link were calculated on the basis of the cost of one additional 'average' train. New 'raw' tariffs were then produced again using the approach adopted for the original Capacity Charge calibration.

However, Network Rail quickly rejected the idea of differentiating tariffs by time and geography. They considered that this was not consistent with the ORR's objective for charges to be "practical, cost effective, comprehensible and objective in function" (Network Rail, 2013c, p11). Particular concern was expressed about the associated billing issues and additional complexity of moving to charges differing by time and geography. Instead Network Rail (2012c) suggested a move from Service Groups to Service Codes which represents a four-fold increase in the number of individual tariffs but did not require fundamental changes to the billing system. Network Rail felt that this provided an opportunity to give "sharper price signals and may incentivise the use of route sections where capacity is more plentiful" (Network Rail 2012c, p14). Network Rail also suggested that where Service Codes were predominantly peak or off-peak services this would address some of the concerns about a lack of time differentiation (Network Rail, 2013c).

Table 4.5 compares the number of Service Codes with the number of Service Groups for a sample of passenger train operators. It can be seen that the ratio between the two types varies widely between operators. Generally, the operators with the greater geographical spread (e.g. First Scot Rail Ltd and Northern Rail Ltd) have a much higher ratio than the main-line operators (e.g. ECML Company Ltd and West Coast Trains). In some cases the move from Service Group to Service Code tariffs will therefore

[17] The reason why CUI can exceed 100% is explained in Chapter Three.

[18] From 121,194 to 88,763 observations (ICL, 2013, p18)

have a much greater impact than for other cases. Although providing greater granularity the move also still means that services with the same code pay the same rate per mile no matter what the time of day is, the location or the direction of travel.

**Table 4.5** Comparison of Service Code and Service Group Numbers for a Sample of Passenger Operators (data Network Rail 2013c, 2008b, analysis by author).

| **Train Operator** | **Service Codes (SC)** | **Service Groups (SG)** | **Ratio (SC/SG)** |
|---|---|---|---|
| Arriva Train Wales | 33 | 8 | 4.1 |
| ECML Company Ltd | 7 | 5 | 1.4 |
| First Capital Connect | 17 | 7 | 2.4 |
| First Great Western Ltd | 42 | 14 | 3.0 |
| First Scot Rail Ltd | 47 | 9 | 5.2 |
| Northern Rail Ltd | 86 | 11 | 7.8 |
| Southern Railway Ltd | 43 | 8 | 5.4 |
| West Coast Trains | 8 | 7 | 1.1 |
| Cross Country Trains Ltd | 12 | 2 | 6.0 |

A weekend discount was again produced by comparing weekday service code tariffs with Saturday and Sunday tariffs weighted by train miles. Arup’s Report (2013, pp33-35) reveals that the average Saturday adjustment was 24.80% lower and the average Sunday adjustment was 42.38% lower. However, Arup noted that traffic was less on Sundays but there were a greater number of possessions for maintenance and renewal. There was therefore not the desire to encourage more traffic on Sundays. They also noted that there was little evidence available that quantified the impact on demand of a significantly lower Sunday tariff. For these reasons a combined average Weekend discount (using weighted averages) was again produced. Following discussions with Network Rail this was rounded up to 33% (compared to 25% for the original charge).

The Freight Flexibility discount and single flat rate were also reconsidered. Arup (2013) noted that Network Rail had much greater flexibility in the timing and routing of freight services which often allowed them to avoid capacity bottle-necks and busy periods. Arup (2013, p35) suggested that “this flexibility is important to the efficient running of the railway, and also important to the efficient allocation of capacity. In light of this flexibility, it is important that a single rate for freight is maintained so that freight operators are not made to pay different rates as a result of Network Rail decisions regarding where to path freight trains". Arup (2013, p37) noted that typically 35% of freight trains are in the long term timetable and the vast majority can be flexed by plus or minus 30 minutes. The remaining 65% of freight trains are planned at less notice and have no restrictions on the level of flex that can be used. Arup (2013, p39) calculated a discount based on the levels of contractual flexibility and proportion of freight services of 21.4%. Following discussions with Network Rail this was rounded up to a 25% discount (compared to the 10% freight flexibility discount applied previously).

Finally as discussed earlier, in the original calibration a de-minimis threshold had been introduced. For the CP5 recalibration exercise the decision was taken not to retain the de-minimis threshold. Arup (2013, p40) tested the impact of retaining a de-minimis threshold (by setting the lowest 10% to zero and recalculating) and found the impact on the calculated tariffs was marginal (average Passenger TOC tariffs decreased by 0.2%, Freight Tariffs decreased by 0.4% and Open Access Tariffs decreased by 0.001%).

### 4.4.5 Financial Implications of the CP5 Capacity Charge Tariffs

Table 4.6 shows the comparison of draft average CP5 tariffs with average CP4 tariffs included by Arup (2013) in their report on the recalibration exercise. The reference to payment rates concerns the Schedule 8 rates (e.g. 'Recalibrated Tariffs (CP4 payment rates)' refers to the tariffs using the new 2013 recalibrated relationships but with the CP4 Schedule 8 rates).

The bottom row of the table shows that for all three categories of traffic the draft Schedule 8 rates contributed a significant part of the total increase in the draft average Capacity Charge tariffs. In the case of the franchised passenger TOCs this represented 83% of the total expected increase. For the other two train types it represented approxaimately half of the overall increase.

Arup (2013, pp45-46) also calculated that changes in the lateness ratio[19] and the infrastructure fault rate between 2002/3 and 2011/12 together

[19] As described earlier in the Chapter, Schedule 8 is based on minutes lateness. An adjustment representing the ratio between delays en-route and lateness therefore needs to be applied to calculate the tariffs.

produced  34% of the overall increase in the tariff values. They also note that overall traffic levels have risen by 13% between 2004/5 and 2011/12.

Arup also looked specifically at the increase in the draft CP5 average freight tariffs. They suggested that the change in the use of the rail network by freight traffic could also have contributed in part to the substantial increase. In particular, the increasing level of inter-modal traffic which characteristically uses more congested parts of the network was particularly referred to by Arup (2013).

**Table 4.6** Comparison of Draft CP5 Average Capacity Charge Tariffs with CP4 Tariffs (2012/13 prices) (Adapted from Arup, 2013, p42)

| | **TOC average** | **Open Access** | **Freight** |
|---|---|---|---|
| Recalibrated CP5 Tariffs (CP5 payment rates per mile) | £1.19 | £3.59 | £0.86 |
| Recalibrated CP5 Tariffs (CP4 payment rates per mile) | £0.59 | £2.07 | £0.47 |
| CP 4 Tariffs (rate per mile) | £0.47 | £0.38 | £0.18 |
| CP5 Increase % | 153% | 846% | 378% |
| CP5 Increase % (excluding CP5 payment rate increase) | 26% | 446% | 160% |
| CP5 Payment Rate impact as % of total increase. | 83% | 47% | 58% |

The very considerable increase in open access tariffs was also believed in part to be due to changes in traffic patterns. Arup (2013, p50) note that traffic on the core ECML route[20] had increased by 22% since 2000, calculating that this factor contributed 17% of the increase. In addition, changes in the fault rate and the Delay:Lateness ratio were found to have contributed 21% of the absolute increase.

In summary, Arup (2013) concluded that the very dramatic rise in expected tariff rates were due to a combination of increases in the Schedule 8

[20] Operated by the two Open Access operators Hull Trains and Grand Central Railways.

payment rates, changes in the lateness ratio and Network Rail's fault percentage and the volume and pattern of traffic on the rail network.

### 4.4.6 The Implementation of the Capacity Charge for CP5.

The ORR published its draft determination on Network Rail's funding and outputs for CP5 in June 2013 (ORR, 2013b). In this it revealed that it had decided against implementing the recalibrated CP5 tariff rates. This was due to the level of the expected significant increases, although the ORR believed that the work carried out for the CP5 recalibration "appears to have been carried out well and to be robust" (ORR, 2013b, p492). Instead, the ORR were minded to approve Capacity Charge tariffs based on the CP4 ones up-rated to account for inflation or to implement an alternative proposal which had been brought forward by freight operators.

The proposal brought forward by the Rail Freight Operators Association (RFOA) was to review actual traffic mileage against benchmarked traffic mileage on a periodic basis. A charge would then be payable if the actual mileage exceeded the benchmarked figure. ORR noted that the expected payments to Network Rail would be substantially less than the Capacity Charge as expected revenue would be close to zero. However, any shortfall in Network Rail's projected variable access charge revenue would be offset through alternative mechanisms.

The ORR noted that "such an approach would allow Network Rail to recover its changes to Schedule 8 costs associated with traffic diverging from the forecast" but "it would be a blunter incentive than the capacity charge because it would apply to all freight operators on an equivalent basis, irrespective of the identity of the operator that had made particular service changes". (ORR, 2013b, p492). The ORR also recognised that setting the charge rates below the calculated increase in marginal costs could dis-incentivise Network Rail from accommodating more traffic on the network. However, their view was that a separate mechanism 'The Volume Incentive Charge' would offset any effect and any loss in revenue would be accounted for by a consequent increase in the Fixed Track Access Charges for the franchised passenger operators.

The Volume Incentive Charge is a mechanism that also encourages Network Rail to accommodate more traffic on the network. The aim is to allow the Infrastructure Owner to share in some of the benefits that operators will gain from running greater than expected additional traffic. In the ORR's final determination, it stated that "the volume incentive should encourage Network

Rail to think about the provision of network capacity to its customers in a more commercial way. This involves making trade-offs when deciding whether to meet unexpected demand" (ORR, 2013c, p725) For CP5 the base-line for the incentive has been set at expected growth, with symmetric incentive rates giving the incentive an expected value of zero.

Table 4.7 shows the value of the Volume Incentive Charge for CP5 compared with CP4. The incentive to Network Rail is based on additional mileage, farebox revenue (for passenger traffic) and load (for freight traffic). It can be seen that the potential incentive has been increased fairly substantially. The floor and ceiling of the charge has also been changed to a limit of plus and minus £300 million respectively (ORR 2013c, p731).

**Table 4.7** Volume Incentive Rates Published in ORR's Final CP5 Determination (source ORR 2013c, p736, adapted by the author)

| | **Final CP5 value (2012/13 prices)** | **CP4 value (2012/13 prices)** |
|---|---|---|
| Per additional franchised train mile | 139p | 84p |
| % of additional farebox revenue | 2.5% | 1.5% |
| Per additional freight train mile | 281p | 136p |
| Per additional freight 1,000 gross tonne mile | 239p | 122p |

The ORR published its final determination on Network Rail's funding and outputs for CP5 in October 2013 (ORR, 2013c). In this it revealed that in light of further industry engagement and consultation it had reviewed its position on the Capacity Charge. The ORR explained that it believed that CP5 Capacity Charge rates should be linked to CP5 Schedule 8 rates because otherwise the "financial disincentives for Network Rail to accommodate additional demand on some routes might result in less efficient use of capacity" (ORR, 2013c, p591). The ORR noted that the Schedule 8 rates for Network Rail have not been updated since 2005 apart from to account for inflation. The rates for passenger trains were increasing by on average 68% which can be explained by large increases in passenger numbers, above inflation increases in fares on some services and updated evidence on how passenger demand responds to increases in journey time (ORR 2013c, p768).

However, it is necessary to comment on the final point. The possibility that passenger sensitivity to increased journey time from disruption is considerably greater than previously thought is curious. Firstly, this appears contrary to Arup (2013, P43) noting that since 2005/6 both CRRD per train mile and primary delay incidents have shown a decreasing trend. In other words sensitivity is greater than previously assumed despite a declining impact. Secondly, the National PPM level as shown in Figure 4.1 of this thesis (P64) has shown steady improvement since the impact of the Hatfield rail crash in the year 2000. A more logical conclusion would be that sensitivity to disruption is actually lower than previously thought as the number of significant incidents declines and passenger face fewer delays (or CRRD) en route. However rather than reflecting an actual change in sensitivity, the impact on the new Schedule 8 rates may instead reflect the output of more studies using more relevant data (ORR, 2013c, P768). In any case, this is a minor point for this thesis, since it is assumed that Schedule 8 is always used to provide the monetary element on any Capacity Charge.

The ORR concluded in its final determination for CP5 that franchised passenger operators would indeed pay the new CP5 Capacity Charge tariff for both existing and new services. Since, franchised operators are protected from any increases in charges for existing services by the Government and could factor any charges for new services into their commercial agreements; the ORR did not consider that there was a “need to mitigate the impact of the charge for them”. (ORR, 2013c, p591).

However, the ORR ruled that existing open access operators would pay CP4 rates for existing services and only CP5 rates for new services. This was because unlike franchised operators they received no protection from the significant increase in the Capacity Charge tariffs in CP5. In making this judgement the ORR were mindful of their statutory duties “to promote the use of the railway network, to protect the interests of users of railway services and to promote competition in the provision of railway services” (ORR, 2013c, p592). In addition, the ORR ruled that new open access operators would pay CP4 rates for services below a threshold set to give similar treatment to existing operators and only CP5 rates above that threshold. This approach was to ensure that operators were being treated in a consistent manner as required by European law and the ORR’ s statutory duties (ORR, 2013c).

For freight operators, the ORR ruled that they would pay a weekday tariff of £0.13 per train mile (i.e. less than the CP4 tariff shown in Table 4.6 with the 25% freight discount applied). At the end of each year there would be a reconciliation based on three commodity groups (coal and biomass, inter-modal and other). The reconciliation would use a base-line of the 2012/13 mileage for each commodity group. The difference between the revenue Network Rail would have received if full CP5 rates were applied to the actual traffic levels for each commodity group above its baseline and the actual revenue received would then be calculated. Any excess would then be apportioned to freight operators by reference to their mileage for the respective commodity groups (ORR, 2013c, pp591-592). The ORR note that if mileage was less than the 2012/13 level the reconciliation amount would be zero (ORR, 2013c). The ORR felt that that this approach would mitigate the significant impact of the calculated CP5 tariff increases for freight operators but incentivise Network Rail to accommodate additional demand. The ORR felt that “it is appropriate to disaggregate the cost reconciliations across three commodity groupings because this improves the incentives for Network Rail to accommodate additional demand” (ORR, 2013c, p592)

The ORR made it clear in their final determination (ORR, 2013c) that the arrangements would only apply for CP5. Whilst recognising the work that Arup and ICL had undertaken and the contribution that the industry had made it was appreciated that this was constrained by short-timescales. The objective is to ensure that a more robust mechanism is in place for CP6 (2020 to 2025).

It is necessary to make some comment about the ORR decision to approve three different approaches to the tariff charge. It is clear that this decision was due to the projected increase in the charge. The reasons given by Arup for this were discussed above. There is also the issue of the impact of the actual methodology and the measure of capacity utilisation (i.e. CUI) on the level of the charge. This will be investigated and discussed as part of this thesis. It is believed though that the application of three different approaches may produce confused and potentially perverse signals. For example, the greater compensation received from new franchised passenger operators services compared to existing open access operators could see the latter ‘squeezed’ away from attractive slots. This may run counter to the policy that the efficient use of the network should consider all calls on the use of capacity at the first instance equally. Nonetheless, it is recognised that the ORR's actions in the case of open access and freight were a recognition that implementation of the calculated tariffs would have a significant impact on their business.

## 4.5 Summary

This chapter has described the real life issues surrounding the performance of the rail network in Britain. It has been seen that although a number of factors affect performance; the volume of traffic and the complexity of the timetable are key issues.

This chapter has also outlined the development and implementation of the Capacity Charge in Britain. The use of the relationship between capacity utilisation and performance has been described. The decision to implement a charge that is not based on disaggregated tariffs by time or geographical location has been highlighted.

The Capacity Charge in CP4 has been reviewed and it has been concluded that it has not been particularly successful in meeting its objectives. The evidence suggests that the specific charge has over-recovered the costs of congestion. However, Network Rail has also stated that the charge has not always recovered the additional performance payments associated with additional traffic. It is therefore unclear whether the Capacity Charge will have always incentivised Network Rail to accommodate more traffic on the network. Secondly, it does not appear that the charge has been entirely successful in its objective of providing price signals to encourage more efficient use of capacity.

The recalibration of the charge for CP5 has produced significant increases in the calculated tariffs which the ORR has felt obliged to mitigate for open access and freight operators. One important aspect of the CP5 recalibration has been a four-fold increase in the granularity of the tariffs from CP4. However, this has still not led to a clear differentiation by time and geographic location, with complexity and transaction costs again being important reasons behind this decision.

It is therefore clear that the Capacity Charge has not been working entirely as intended. Although, there have been an increase in the granularity of the charge for CP5 through the adoption of Service Codes and the approach adopted for the recalibration is considered "robust", capacity utilisation has continued to be measured using CUI and the methodology for calculating the tariffs has remained essentially the same.

The next chapter describes the methodology that will be used to conduct a new analysis of the relationship between capacity utilisation and performance, with the results being used to assess the implications for the pricing of congested rail networks.

# Chapter 5
# Methodology

## 5.1 Introduction

This chapter outlines the methodology used to carry out a new regression analysis with the variables described in Chapter Three. As discussed in the previous chapter the comprehensive approach used in the 2013 recalibration of the Capacity Charge provides an excellent framework for this analysis and has therefore been adopted. Any divergences from this approach are clearly identified. One advantage of using the same approach is that comparisons can be made with the conclusions obtained from this national exercise.

The second part of this chapter explains some general principles about the production of the data set used in the analysis. Details of the actual data set are provided in the next chapter.

The final part of this chapter explains how the results of the regression analysis were applied to the question of the pricing of congested rail networks.

## 5.2 The Regression Analysis

### 5.2.1 General Principles

The next chapter describes the creation of the actual data set for the analysis. However, in order to explain the methodology adopted it is necessary to understand some general principles.

The description of the Capacity Charge methodology by Arup (2013) and ICL (2013) were used as the basis for the approach adopted for this thesis. However, standard econometric text books by Dougherty (2011), Kennedy (2008) and Wooldridge (2002) were also consulted.

The analysis was undertaken using the EViews software package.

The data is divided into a number of geographic sections (i.e. cross-sectional data) and time bands (i.e. time series data). This matrix of data lends itself to the use of a ‘Panel Data’ approach which was therefore adopted. Two different panel data sets were created. A larger one was used to test the sectional explanatory variables and the smaller one the area explanatory variables. Both data sets are balanced i.e. each geographic location has the same number of time-bands. The creation and contents of the data sets are described in detail in the next chapter.

A number of different functional forms were tested. These were taken from the 2013 recalibration of the Capacity Charge. However, since they are a mixture of linear and non-linear forms it was necessary to carry out a transformation of the data set. The approach adopted was the Box-Cox data transformation technique (Dougherty, 2011)

Following the technique used for the recalibration exercise, 'fixed effects' and 'random effects' approaches were compared as were 'one-way' and 'two-way' approaches. This gives a total of four different approaches[21]. Arup (2013) note that this approach was adopted to account for any omitted variable bias and any confounding (i.e. an omitted variable that correlates directly with both the dependent and explanatory variable).

'Fixed effects' and 'random effects' were compared using a Hausman test (Kennedy, 2008). The choice between one-way and two-way; functional form and explanatory variable was then made using standard measures of 'success' which will be described later in the Chapter.

Finally, any evidence of auto-correlation and heteroskedasticity were accounted for using standard techniques.

### 5.2.2 Functional Form

For consistency, the functional forms used in the recalibration of the Capacity Charge in Britain (Arup, 2013, p20) were used in the regression analysis undertaken for this thesis. These are:-

Linear $D_{it} = A_i + \beta C_{it}$ (20)

Exponential $D_{it} = A_i * \exp(\beta C_{it})$ (21)

Quadratic $D_{it} = A_i + \beta C_{it}^2$ (22)

Second Order Approx. $D_{it} = A_i + \beta_1 C_{it} + \beta_2 C_{it}^2$ (23)

Where:
$D_{it}$ is the reactionary delay per train mile on section i in time period t;
$A_i$ is a section (area) specific constant;

[21] i.e. fixed effects with a one-way approach; fixed effects with a two-way approach; random effects with a one-way approach and random effects with a two-way approach.

β is the coefficient of the capacity utilisation which in the original calibration varied by route (Gibson, S. Cooper, G, and Ball. B. 2002) but in the re-calibration of the Capacity Charge is a network-wide value (Arup, 2013) and $C_{it}$ is the calculated capacity utilisation percentage for section *i* and time period *t*.

As with the 2013 recalibration of the Capacity Charge the Second Order Approximation used both a linear and a logarithmic form. Therefore, a total of five different equations were used for the analysis.

Three of the equations were linear in form (i.e. Linear, Quadratic and the linear version of the Second Order Approximation specification) and two were non-linear (i.e. Exponential and the logarithmic version of the Second Order Approximation specification).

Since a number of cells in the data sets had a CRRD value of 0, it was necessary to make an adjustment due to the use of logarithms. The standard approach of adding 1 to the level of reactionary delay (i.e. giving a dependent variable of (CRRD+1) / Train Miles), which had been used in the 2013 recalibration (ICL, 2013, p12), was therefore used.

As noted previously, the Exponential relationship was adopted by both the original calibration of the Capacity Charge (Gibson, S., Cooper, G. and Ball, B. 2002) and the subsequent re-calibration (Arup, 2013). It is also consistent with the findings of the literature review discussed in Chapter Three.

### 5.2.3. The Box-Cox Transformation of the Data.

The mixture of linear and logarithmic based dependent variables also meant that a transformation of the data set is necessary to allow the different functional forms to be compared. The standard Box-Cox approach was used with the method outlined by Dougherty (2011, pp205-207) being applied. This has the advantage over other methods of Box-Cox transformations that in this case a linear regression approach can be used.

In this procedure the observations are scaled on the dependent variable (Y) so that the residual sums of squares in the linear and logarithmic models are rendered comparable. Dougherty (2011) notes that the procedure has the following steps:-

1. The geometric mean of the values of Y in the data base is calculated. This equals the exponential of the mean of log Y.

2. Observations are scaled on Y by dividing by this figure (i.e. $Y_i$* = $Y_i$ / geometric mean of Y) where Y* is the scaled value in observation $_i$.
3. The linear models are then regressed using Y* as the dependent variable and the logarithmic model use log Y* as the dependent variable.

The residual sums of squares obtained from this approach were then used to decide between the linear and logarithmic functional forms. As noted by Dougherty (2011) however it is then necessary to revert to the original non-transformed data to complete the regression analysis.

### 5.2.4 Fixed Effects and Random Effects

The next step in the regression analyses is to determine whether ‘random’ or ‘fixed’ effects provide the most appropriate means to account for the impact of any omitted variables (or unobserved effects) on the strength of the derived relationships. It is also necessary to allow for any possible confounding (i.e. an omitted variable that correlates directly with both the dependent and explanatory variable). Wooldridge (2002, p252) explains that “in modern econometric parlance” “’random effect’ is synonymous with zero correlation between the observed explanatory variables and the unobserved effect” and the term ‘fixed effect’ means one is allowing for arbitrary correlation between the unobserved effect ... and the observed explanatory variables”.

Kennedy (2008, pp283-286) talks about the two approaches in terms of omitted variable bias. He notes that if the collective influence of any unmeasured omitted variable is uncorrelated with the included explanatory variables then omitting them will not lead to any bias in the regression model. These omitted variables can therefore be included in the error term without causing any bias and random effects is used. However, if there is correlation between the omitted variables and the included explanatory variables then they need to be included in the model since omitting them causes bias. The fixed effects approach does this by including a dummy variable for each cross-sectional unit.

In the case of the relationship between capacity utilisation and the level of reactionary delay, a random effects approach would therefore imply that any unexplained delay was due to other variables that had no link to the level of traffic on the infrastructure. This impact is captured in the EViews software within the *‘A’* constant which therefore incorporates a random error term. In contrast a Fixed effects approach implies that there is some kind of link between variables that had not been modelled and the level of traffic on the infrastructure. The dummy variables for each cross-section are included in the '*A*' constant in the specification.

The Hausman Test is considered the standard test for choosing between fixed and random effects. This compares a null hypothesis that both approaches are equally consistent with an alternate hypothesis that only a fixed effects is appropriate due to the potential for bias in with the random effects approach. In the event that the null hypothesis is not rejected, the recommendation is that random Effects is adopted due to the inefficiency produced in fixed effects by the need to create a number of dummy variables and the consequent loss of degrees of freedom.

However, Dougherty (2011, p525) makes the very clear point that if the sample used in the regression is non-random then a fixed effects approach should be used. It can be argued that although the findings from this analysis are intended to be transferrable, the data is not random. The areas were chosen due to known congestion issues and are therefore not a random sample of the British rail network as a whole. Secondly, the areas were chosen due to their specific characteristics and are therefore not a random sample of congested parts of the rail network. This suggests therefore that a fixed effects approach should be adopted. However, a Hausman Test will still be undertaken to establish its results.

Both the original Capacity Charge work and the subsequent recalibration adopted a fixed Effects approach (Gibson, S., Cooper, G. and Ball, B., 2002 and Arup, 2013).

### 5.2.5 'One-way' and 'Two-way' models

As previously noted, the balanced panel data sets used for this analysis are divided by infrastructure section (cross-section) and time band (time series). The next step is to determine whether a one-way or two-way model approach is the most appropriate.

In the one-way models that will be used in this analysis, the assumption is that the variation in infrastructure between different geographic sections also has an impact on the relationship between capacity utilisation and reactionary delay. A one-way model therefore produces a constant that varies by section (or area). The equations shown in section 5.2.2 of this thesis therefore represent one-way models.

The idea that local variations in infrastructure should have an additional influence on reactionary delay appears logical. For example, as described in

Chapter Three although the planning headways and junction margins that form the basis of the capacity utilisation calculations are based on the capability of the infrastructure these are by necessity simplifications of local conditions. The use of a section specific constant is also likely to capture the influence that adjacent links and nodes will have on reactionary delay. The latter is also investigated in this analysis via the capacity utilisation variables which include junction moves (XCUI and XHET) and the ‘Section Before’ and ‘Section After’ variables. Significantly, a one-way model approach is the one that was adopted for the original calibration and subsequent recalibration of the Capacity Charge (Gibson, S., Cooper, G. and Ball, B., 2002 and Arup, 2013).

Two-way models produce constants for both cross-sectional variation and time variation. The concept that there should be variation between the levels of reactionary delay due to the influence of the specific time of day does not appear to have much basis in logic. This is because the only key variation on the rail network between different time periods is the level of capacity utilisation. This is obviously already captured by the explanatory variable. One possibility though is that time periods following ‘peak’ periods will experience greater levels of reactionary delay due to the residual effects of the high capacity utilisation previously. However, modelling this potential effect through a two-way model assumes that the peak periods in the sample network all correspond to the same hourly time periods. The inclusion of the ‘Time Period Before’ variable in the analysis is seen as a more effective means of examining this potential effect.

Alternative models that will not be investigated as part of this analysis is firstly, one that assumes no cross-sectional or time period variation and secondly, a one-way model that produces a time period constant. It seems extremely unlikely that reactionary delay in the whole sample network could be explained by the calculated capacity utilisation alone. The variation in infrastructure described earlier in this chapter supports this view. The use of one-way or two-way models allows a finer level of detail to be modelled. As noted whilst section specific constants appear to be logical, time period specific constants appear less logical. Investigating the value of a one-way model with time specific constants therefore appears to have little merit.

### 5.2.6 The Decision Criteria

In order to compare the different explanatory variables and the different functional forms it is clearly necessary to adopt some form of decision criteria. The original capacity charge work used the t-statistic to determine the most appropriate functional form (Gibson, S., Cooper, G. and Ball, B. , 2002). Although, the 2013 recalibration report refers to a number of different criteria; the key measure adopted was the R-squared value (Arup, 2013). Both the t-statistic and the R-squared value are standard methods of determining the strength of the relationship between the dependent variable and the explanatory variables.

The t-statistic is used to test the likelihood that a parameter value is equal to zero. In other words, there is not a significant relationship between the dependent and explanatory variables. The size of the t-statistic values in the output of the various regression analyses will be compared to determine the significance of the different capacity utilisation and ‘other’ explanatory variables. This is done by comparing the value of the statistic against a standard value.

However, in the case of the Second Order Approximation functional form the t-statistic is inappropriate to determine whether the specification is correct as the value is ‘shared’ between both capacity variables. For this reason the F-test of Joint Significance has been used to determine whether the functional form is suitable. The methodology described by Dougherty (2011, pp180-182) has been employed. In this :

- A regression is first run for the data set using the constant alone. This is followed by a regression for the full specification.
- The residual sums of the squares (RSS) are taken from both sets of results.
- The reduction in RSS is then calculated as the RSS for the constant alone minus the RSS for the full specification. $RSS_1 – RSS_2$ is then divided by the cost in the degrees of freedom (the number of additional parameters estimated). The result is the numerator for the calculation .
- The denominator is $RSS_2$ divided by (the number of observations minus the number of degrees of freedom).
- Dividing the numerator by the denominator produces the F-value which can then be compared with tables of significant values.

The R-squared value is often described in terms of ‘goodness-of-fit’ i.e. how closely the modelled relationship matches the actual data points in the data set. Kennedy (2008, p13) explains the measure as “the proportion of variation in the dependent variable ‘explained’ by variation in the (explanatory) variables”. Due to the presence of more than one explanatory variable in some cases, it is necessary to adopt the alternative ‘adjusted R-squared’ measure which accounts for the effects of this in the results. Kennedy (2008, p26) also notes that “in dealing with time-series data, very high $R^2$s are not unusual, because of common trends” but “for cross-sectional data, typical $R^2$s are not nearly so high”. Since, the data sets are a combination of the two types of data high adjusted R-squared values are not necessarily a pre-requisite for determining that a particular modelled relationship is acceptable. It is noted that the $R^2$s values obtained in the 2013 re-calibration were extremely low (Arup, 2013). Kennedy (2008, p89) further notes that searching for a high $R^2$ value “runs the real danger of finding through perseverance, an equation that fits the data well but is incorrect because it captures accidental features of the particular data set at hand ...rather than the true underlying relationship”.

### 5.2.7 Autocorrelation and Heteroskedasticity

It is then necessary to test for the presence of autocorrelation and heteroskedasticity and if necessary make the appropriate adjustments to the regression outputs.

Autocorrelation, or serial correlation, is where there is a correlation between the error terms of different observations. Dougherty (2011, p429) explains that autocorrelation normally occurs only in regression analysis using time series data and is generally “persistence of the effects of excluded variables”. It is necessary to check for its existence in the generated models as its presence could lead to inefficient results and the potential for erroneous conclusions.

Heteroskedasticity is the phenomenon where the size of the error term does not exhibit constant variance. A common cause of this is an increasing difference between actual observations and the ‘fitted line’ produced by the regression process as the size of the units measured increases. For this analysis, this would mean a greater difference between observations and the fitted line as the measured level of capacity utilisation increases. Logically there is a possibility that this could occur in this analysis. This is because as previously discussed increased capacity utilisation suggests a greater likelihood of reactionary delay propagation but the primary incidents themselves are essentially random events. Once again it is necessary to check for and account for heteroskedasticity in the results due to the possible inefficiency and the likelihood of erroneous conclusions.

The EViews software allows the possibility of both autocorrelation and heteroskedasticity to be accounted for in a number of ways. A White Heteroskedasticity Consistent Covariance Matrix was chosen as this adjustment is designed to account for autocorrelation and heteroskedasticity of unknown form.

### 5.2.8 Instrumental Variables

An alternative problem is where a variable is conceptually different from the true explanatory variable in the relationship. If the explanatory variables used in the analysis have a random component that is not distributed independently of the error term, i.e. there is a link between the two, the results of the regression analysis will lead to biased estimates of the parameters. In order to test for this problem an Instrumental Variable (or IV) approach is commonly used. Dougherty (2011, p316) explains that "essentially IV consists of semi-replacing a defective explanatory variable with one that is not correlated with the (error) term". In the IV approach, another variable is used which is correlated with the capacity utilisation variable but not the error term.

The method used in the recalibration exercise to create the alternate variable (or instrument) was the Durbin Rank method as referred to by Kennedy (2008, p142). In this method the explanatory variable is ranked (i.e. the cell with the highest capacity utilisation is given the highest rank) The use of this ranking method ensures a strong correlation between the instrument and the explanatory variable . This rank is then used as the instrument. In order to check that the instruments are suitable, their strength is checked by calculating the correlation between the rank and the measured capacity utilisation (i.e. the explanatory variable). Assuming that the instruments are suitable they are then used in a new 'two-stage' regression analyses which is used to identify any problems with measurement bias in the explanatory variables.

However, it is not clear that the Durbin Rank method produces an instrument that is completely uncorrelated with the error term. Any significant error (or defectiveness) in the capacity utilisation variables could affect their ranking and means the instrument is affected by errors Kennedy acknowledges the possibility of this with this methodology and notes it cannot be tested for with only one instrument (p144).For this reason this approach has not been pursued. Indeed it has proved impossible to identify an instrument approach that appears acceptable. As Kennedy notes (2008, p143) "regardless of how cogently the validity of the instrument, disputes can arise concerning the need for instruments, the validity of the instruments and the interpretation of the IV coefficient estimates". Instead, it has been assumed that the explanatory variables described in Chapter Three are acceptable on the basis that they are well grounded in accepted theory.

### 5.2.9 Multiple Regressions

The introduction of the 'other' explanatory variables adds an additional level of complexity to the regression analyses.

As described in Chapter Three the objective of introducing 'other' variables to the regression equations is to establish whether a more effective explanation of the causes of reactionary delay could be achieved. However, rather than repeating the entire regression process; this stage is carried out once the most effective functional form and model form (i.e. one-way or two-way and 'random' effects or 'fixed' effects) have been established.

The regression analysis is then repeated for each of the capacity utilisation measures but this time with the addition of the 'other' variables to the equations. The t-statistic test will be used to establish whether these other variables are significant when combined with the capacity utilisation variables. Once any significant 'other' variables have been identified the impact of using multiple explanatory equations on the 'decision criteria' will be compared with the original results. Additionally, each of the 'other' variables that are not associated with capacity utilisation (i.e. not the 'Time Before', 'Section Before' and 'Section After' variables) will be examined in individual  regression analyses to establish whether non-capacity utilisation measures provide more effective explanations of the causes of reactionary delay. The results will be used to decide whether a single capacity utilisation measure is the most appropriate means of predicting reactionary delay and thus forming the basis for a congestion charge.

Of course a significant danger in multiple regression analyses is the risk of colinearity i.e. where the correlation between two explanatory variables makes it difficult or impossible for the model to predict the relationship with the dependent variable The EViews software rejects any equations that it identifies as suffering from perfect colinearity.  Additionally, the correlation between the capacity utilisation explanatory variables and any 'other' variables that have been identified as significant will be calculated. The results of this comparison also aids the decision about whether it is worth the additional complexity of adding more explanatory variables to the regression equations.

## 5.3 Creation of The Data Set

### 5.3.1 Overview

The next chapter details the specific data sets used to undertake the analyses described in this thesis. However, the methodology used to prepare them is described in this section.

The decision was taken to focus on a timetable for two parts of one route (the ECML). This would allow an investigation of the relationship between capacity utilisation and performance to be carried out in some detail. As described earlier in this thesis, the transferability of any conclusions from this sample network would then be considered.

### 5.3.2 The Timetable

The type of timetable used to calculate capacity utilisation is an important issue that needs to be discussed. As previously highlighted access to the rail network for services is via inclusion in a timetable. However, trains are planned right up to the day of operation. Almost all passenger trains are included in the so-called Permanent Timetable which is completed approximately 6 months before each timetable change date. This is to allow for publication of the passenger timetable. There are however some later revisions to passenger services but these are principally re-timings for engineering work.

In contrast, as noted by Arup (2013, p37), 65% of freight services are planned later than the Permanent Timetable completion date. This is due to the flexible nature of the freight business. This timetable ‘fluidity’ has a bearing on the capacity utilisation calculations since the results will potentially vary considerably depending on the point in the timetable process that is selected.

For the recalibration, Arup (2013) chose to effectively use the day of operation. This means that all traffic which operates on a given day (including all freight traffic) is included in the capacity calculations. However, the varying nature of the timetable means that each day will potentially differ from the next. To address this issue, Arup (2013, p8) chose one ‘representative’ day to represent  weekdays, one to represent Saturdays and one to represent Sundays. Although, accurate calculations can be carried out for those days there is clearly a two-fold risk in adopting this approach. Firstly, one-off trains that do not run on any other day in the timetable may run on the day in question. Secondly, there may be trains that operate on the vast majority of days but not on those selected. Both issues will distort the results of the capacity calculations.

For this analysis an alternative approach was chosen. The Permanent Timetable itself was used as the basis for the capacity utilisation calculations. As described in the next section, freight paths in the timetable that did not actually run or only did so very occasionally were excluded from the data set. This approach avoids the issues with the ‘representative’ day approach described above but excludes new paths and changes made to existing paths. Both approaches therefore have advantages and disadvantages. Each approach is a compromise. In the end the main factor in the decision to use the Permanent Timetable was that it could easily be obtained from Network Rail.

Finally, it needs to be noted that a weekday timetable was chosen for the analysis. This is due to the greater volume of traffic during the week than at the weekend and the likelihood of there being congestion is consequently increased.

### 5.3.3 Source Data

Timetable and Performance data were supplied by Network Rail. Both sets of data came in two parts.

The timetable data supplied by Network Rail took the following forms:-

- Timetable reports for each key timing point in the data set. These list each service in time order and as appropriate by line, crossing move and designated platform. These reports therefore give detailed information on the timetable at specific locations.
- Timing schedules for each planned service in the sample data set. These schedules detail the planned times, lines, allowances and as appropriate station stops for each timing point for a train’s entire journey. These reports therefore give detailed information on the timetable for specific train paths.

The timetable reports were used to allocate trains to specific geographic sections and time bands. They were also used to identify the size of gaps between consecutive trains as well as the basis for ‘graphing’ the relevant trains in each cell. They were therefore used as the main source of information for the calculation of the traffic Intensity, CUI and HET variables. They were also used to assist with the allocation of CRRD to the correct cell.

A very small extract of a weekday timetable report for Stevenage for the December 2008 to May 2009 timetable is reproduced as Table 5.1.

For reasons of space only the Up direction and a very small amount of time is shown. Nonetheless the use of the Up Fast, Up Slow and Down Slow lines (FL, SL and DSL respectively) at different times can be clearly seen. The presence of passing trains (denoted by a 'p') and stopping trains (denoted by the relevant platform number and an arrival and departure time) is also clear. The trains themselves are also identified by their head-codes and their origins and destinations.

**Table 5.1**. Sample Timetable Report for Stevenage (December 2008 to May 2009 Timetable, supplied by Network Rail).

| Train | Origin | Destination | Line | Time | Line | Platform |
|---|---|---|---|---|---|---|
| 1P63 | Peterborough | Kings Cross | SL | a 0800 | | 1 |
| 1P63 | Peterborough | Kings Cross | | d 0801 | SL | 1 |
| 2J21 | Stevenage | Moorgate | | d 0805 | DSL | 4 |
| 1P54 | Peterborough | Kings Cross | FL | p 0805½ | FL | |
| 1R53 | Royston | Kings Cross | SL | a 0808½ | | 1 |
| 1A05 | Leeds | Kings Cross | FL | p 0808½ | FL | |
| 1R53 | Royston | Kings Cross | | d 0809½ | SL | 1 |

The individual timing schedules were used to check the train timing details in individual cells and in particular that the CUI and HET information had been entered correctly. The 'point-to-point' timings were used to produce the 'Average Speed' explanatory variable (Equation 17); whilst the performance and pathing allowances were used to produce the 'Stability' explanatory variable (Equation 15). The schedules also give the Service Code for that particular train and therefore the information to calculate the 'Timetable Complexity' explanatory variable (Equation 16). The timing schedules also give the route of each particular train from their origin. This information was used along with the mileage information in the relevant track diagrams to calculate the 'Average Distance Travelled' variable (Equation 19).

**Table 5.2** Extract of Timing Schedule for 1P63 (December 2008 to May 2009 Timetable, supplied by Network Rail).

| Location | Arrive | Depart | Platform | Line | Perform. Allowance | Pathing Allowance |
|---|---|---|---|---|---|---|
| Sandy | 0739 | 0740½ | | SL | | |
| Biggleswade | 0743½ | 0744 | | SL | | |
| Arlesey | 0748½ | 0749 | | SL | | |
| | | | | | | ½ |
| Hitchin | 0754½ | 0755½ | | SL | | |
| Stevenage | 0800 | 0801 | 1 | SL | | |
| Woolmer Green Jn | 0804½ | | | FL | | |

Table 5.2 reproduces a small extract of the timing information for 1P63, a Peterborough to Kings Cross train, which is the first train shown in Table 5.1. It can be seen that 1P63 uses the Slow Line for most of the journey shown in Table 5.2, except at Woolmer Green Junction where it crosses to the Fast Line. The train stops at all the station locations listed but only Stevenage has a specific platform designated. Finally, half a minutes pathing allowance is allocated to the train after Arlesey.

The timetable information was also used to check the days that each train in the timetable was planned to operate. A number of timetabled trains, typically freight trains, are only scheduled to operate on a limited number of days in the week. For example, a freight train might be planned to only run on Wednesdays. The inclusion of these limited paths will therefore increase the calculated capacity utilisation figures despite them not being planned to operate on the majority of days. However, a single planned service might be the cause of a considerable amount of CRRD when it is present. Once again a compromise is necessary. Any train only planned to operate on single days of the week were excluded from the data set. An important caveat is that some of these trains share core paths with other trains that run on different days of the week. In other words, the actual core path runs on multiple days of the week. Where the core part of the path was relevant to the sample network it was therefore logical to include it in the calculations.

The two types of performance data provided was:-

- A data base containing individual train delay records for the sample timetable and network.
- A data base containing individual train lateness records for each timing point in the sample timetable and network.

The ‘delay’ database was used to produce the dependent variable for the regression analyses. For every train suffering an incident of delay in the timetable the record includes:-

- Headcode of Train affected
- Delay Code.
- Date of Incident.
- Start of section delay occurred.
- End of section delay occurred.
- Incident Serial Number
- Incident description
- Headcode of Train responsible for the delay.
- Delay Minutes suffered.

This information can be used to produce a detailed picture of performance issues on the sample network. For example, the data set includes an incident of 5 minutes reactionary delay incurred by the East Coast train 1A05 between Sandy and Hitchin on 11th February 2009. This was due to following a late running train (1P54) which had lost its path (reactionary delay code YD) due to a power dip in the over-head lines at Templehirst[22].

Once the data set had been filtered to remove non-CRRDs and weekend incidents, the delay records were sorted by geographic section and time band. This was achieved using the start and end sections and the train headcodes. Due to its importance this data allocation phase will be returned to in the following sections. There are also some differences to the approach adopted for the 2013 Capacity Charge recalibration and these also need to be highlighted.

The CRRD delay for each cell was then converted into the dependent variable (i.e. CRRD per train mile) using the relevant train numbers in each section and time-band and the mileage for the section calculated using

[22] Once again this example illustrates the propagation of reactionary delay. Templehirst is 125 miles north of Sandy on the ECML.

railway track diagrams (Trackmaps, 2005). As discussed previously CRRD was increased by 1 before being divided by train miles, so that logarithms could be applied to the non-linear functional forms.

The individual daily lateness records compares the actual time with the planned time for each train in the timetable at every timing point in the sample network. An example of the contents of this data set is shown in Table 5.3. This presents a week's lateness records at Grantham for the East Coast train 1E12 Inverness to London Kings Cross. It can be seen that the train was on time on one of the days (6th February), early on another (4th February) and late by varying degrees on the remaining three days.

The lateness data was used to calculate the 'Average Entry Lateness' variable[23] . It was also used to help identify which Freight trains in the sample timetable to include in the data base. Apart from the freight paths that were planned for only single days in the week, there are other more frequent paths in the Permanent Timetable that never or only occasionally operate. It is important that these are also excluded from the capacity calculations, otherwise the results will suggest a higher utilisation than was actually the case. The lateness records were therefore used for each freight remaining path that remained following the removal of the 'single day' paths. Those that did not actually operate or only operated on less than 5% of days were also excluded.

**Table 5.3** A Week's Lateness Record for 1E12 at Grantham.

| Date | Minutes Late Compared with Planned Time |
|---|---|
| 02 Feb. 2009 | 10.5 |
| 03 Feb. 2009 | 4 |
| 04 Feb. 2009 | - 5.5 |
| 05 Feb. 2009 | 2 |
| 06 Feb. 2009 | 0 |

Another source of data used in the analysis was the Timetable Planning Rules for the relevant timetable (Network Rail, 2009a). As described

[23] Equation (16)

previously, these contain the necessary information on the planning headways and junction margins to be used in the calculations.  Finally as also mentioned previously, track diagrams for the relevant parts of the network were used. These give information on mileages and the layout of the actual infrastructure.

### 5.3.4 Division into Geographic Sections

The nature of the timetable and performance data obtained from Network Rail has implications for how the analyses is conducted. One important issue is the division of the sample network into geographic sections.

The Capacity Charge  used  Constant Traffic Sections (CTS) which are sections within which “train counts are constant i.e. no trains start, terminate, join or leave between CTS ends” (Arup, 2013, p5).

Although, this is technically more accurate in terms of the calculation of capacity utilisation at a local level; it was found with the data supplied by Network Rail that use of CTSs leads to a mismatch between the geographic sections and the performance data. The delay and lateness data is based on significant timing locations which themselves are based on major stations or junctions. This means that one ‘performance’ section could consist of a large number of CTSs. This is something recognised by Arup (2013, p15). Indeed they note that the timing locations used for some groups of train  do not include all the relevant CTS locations. To address this, train times for each CTS location were calculated based on the interpolation of the scheduled times across the intermediate CTSs in proportion to their lengths.

CRRD was then allocated between the CTSs on a pro rata basis. So for example a ‘performance’ section with 100 minutes CRRD and consisting of 10 CTSs would have 10 minutes delay given to each of them. There is an obvious issue with this approach. It implies that there is a uniform spread of delay along the ‘performance’ section. This may hide the influence of the start and end nodes of the section on performance which as noted will be important stations and junctions. It seems more likely that rather than an even spread of delay, CRRD is potentially higher in the vicinity of areas more likely to suffer from congestion.  The use of the CTS approach also implies that traffic levels on adjacent sections will differ but as noted CRRD has been equally distributed resulting in potential problems.

The approach adopted for the 2013 Recalibration may be termed an ‘infrastructure led’ approach as the delay data is ‘made to fit’ the geographic sections. It can also be considered a microscopic level approach due to the level of detail that the CTSs represent. Gille, A., Klemenz, M. and  Siefer, T. (2010) proposed three levels of detail for the modelling of railway infrastructure; namely microscopic, mesoscopic and macroscopic. Microscopic modelling uses a very fine level of detail and is typically used for precise capacity allocation at individual train level. In contrast, macroscopic modelling uses a general level of detail and is typically used for long term 'broad brush' strategic planning. Between the two levels of detail is mesoscopic. The approach used for the creation of the geographic sections for this thesis can be considered a mesoscopic approach.

For this thesis a 'performance led' approach was adopted. In other words the geographic sections matched the sections from the performance data. It was therefore not necessary to divide the CRRD between several different geographic sections. It is however necessary to take into account trains that were timed on only part of the geographic section. The approach adopted can be illustrated using examples from the ECML (the route from which the sample network for the analyses is taken).

Some locations such as Claypole Loop which is located within the Grantham to Newark North Gate section, is relatively straightforward as it is also a mandatory timing point for all traffic so it is simple to account for its presence in the capacity calculations.

Other timing points that are not used by all trains require a different approach. One feature of the ECML Train Planning Rules is that in order to facilitate timetable construction, the times of trains using these 'secondary' timing points are linked to mandatory timing points. This feature is illustrated in Table 5.4.

The table shows the relationship between two different types of train for two different locations on the ECML. The first is for Digswell Junction which in the Up direction is the location where 'combined' traffic on Welwyn Viaduct divides again into Fast and Slow Line traffic. Only trains crossing onto the Slow line are timed at Digswell Junction. Table 5.4 shows that the margin between a train crossing to the Slow line and the next Up Fast train applies to the time that the latter passes its next mandatory timing point i.e. Welwyn Garden City. It can be seen that the second example for Carlton Loop follows exactly the same format. These rules provide a means for establishing how much capacity is used by a given timetable.

**Table 5.4** Example of the Link Between Mandatory and Secondary Timing Points on the ECML (Network Rail, 2009b, p45 and p59).

| Location | First Train | Second Train | Margin (minutes) |
|---|---|---|---|
| Digswell Jn | Up Train crosses to Slow Line | Up Fast passes Welwyn Garden City | 3.5 |
| Carlton Loop | Down Train arrive | Next Down Train passes Newark North Gate | 2 |

An additional element of the methodology that needs to be discussed is the handling of multiple tracks. For the recalibration of the Capacity Charge, the calculated capacity utilisation for each of the tracks was summed and then divided by the number to give an average CUI value (Arup, 2013, p13). For this analysis, Fast and Slow lines were treated as separate geographic sections and the results only combined at the tariff calculation stage[24]. The advantage is that the impact of very high capacity utilisation is not reduced in the averaging process by a much lower utilisation on the adjacent line.

There is however the issue of the allocation of CRRD to lines where trains crossed between parallel lines within a geographic section. This was relatively uncommon in the data set used for the analysis. However, where it was necessary to divide CRRD between Fast and Slow lines; rather than using the pro-rata approach adopted by Arup (2013), the record of which train was responsible for the reactionary delay was used to decide which geographic section the CRRD belonged to.

In the sample network there was one location where the pro-rata allocation of CRRD was necessary. This was at Langley Junction within the Stevenage and Welwyn Garden City Slow Line geographic sections. This is where traffic interacts between the ECML and the Hertford Loop. Langley Junction is only a timing point for traffic using the Hertford Loop. Delays in the data set are listed for both the Stevenage – Woolmer Green Junction and Stevenage – Hertford North (the first station on the loop) sections. CRRD listed for Stevenage – Hertford North was allocated to the Stevenage – Woolmer Green geographic sections on a pro-rata mileage basis. This is because it is not clear from the data set whether the CRRD occurred on the ECML or the Hertford Loop.

[24] As discussed later in the chapter.

One element where there was some agreement between the Capacity Charge recalibration and the analyses carried out for this thesis was the allocation of nodal delay. This type of delay has a single location in the CRRD data set and typically refers to major junctions or stations. A decision has to be made about which adjacent geographic section the delay is allocated to. In the Capacity Charge recalibration the delay was placed in the next section on the basis that “the cause is located immediately ‘downstream’ of the recorded location” (Arup, 2013, p15). For example, at a station a train might suffer reactionary delay waiting for a path onto the ‘downstream’ section.

This appears to be contrary to the method discussed in Chapter Three for the calculation of link and node capacity utilisation. In this case it is assumed that the impact of junction congestion will typically be felt at the end of a link. This was because traffic approaching a junction would be affected by any congestion in terms of having to slow down or stop.

This apparent contradiction can be explained by the fact that traffic at stations and junctions behave differently when they are delayed. Delays at station nodes refers to trains waiting at the station itself for a path. It is therefore appropriate to allocate any nodal delay into the next adjacent section. This was done for this thesis. However, trains stopping at junctions do not stop at the actual location but instead will come to a halt at the signal immediately preceding it. Any junction nodal delay was therefore allocated to the link immediately preceding the node.

### 5.3.5 Division into Time-Bands

In contrast to previous work on the Capacity Charge, hourly time-bands were used in order to maximise the amount of data for the analysis. This allows a more accurate picture to be obtained of how changes in the level of capacity utilisation affect the levels of reactionary delay. An hourly period was considered the smallest practical unit for the analysis (such a view is for example supported by Gibson, S., Cooper, G., and Ball, B.,2002). One reason is that this accounts for the possibility of an hourly repeating timetable capturing each element of this within a single capacity utilisation figure.

There are a number of issues surrounding the methodology that need to be explained:-

- The ‘handling’ of trains straddling time-bands.
- The ‘linking’ of adjacent time-bands via train journey time.

For the Capacity Charge recalibration, trains that straddled time bands (i.e. their ‘entry time to’ and ‘exit time from’ a CTS were in different time bands) were allocated to the time band their median time belonged to. To avoid this additional calculation, trains that straddled time-bands in this analyses were allocated to the time-band they entered the geographic section. Since the CRRD ‘belonging’ to a train is allocated to the same specific cell (i.e. the entirety of the performance impact is placed in the same cell as the entirety of the capacity utilisation) it is not believed there are any issues with this approach.

Curiously though Arup (2013, p13) note that freight traffic timed to wait in loops for other traffic to pass and thus straddling time-bands was excluded from the data set. This was due to the significant increase in capacity utilisation this represented. For this analysis, looped freight traffic was treated as not being on the sample network. The path from the ‘start of the geographic section to the loop’ and the path from ‘the loop to the end of the section’ were treated as separate partial paths as described previously. Where these partial paths were allocated to separate time-bands; CRRD was allocated to the correct portion using the ‘Train Responsible for the Delay’ field in the delay data set to identify where in the timetable the delay occurred.

One final aspect is the need to adjust the time-bands of adjacent geographic sections so that the results can be directly compared. This is necessary due to the effect of distance. For example if Section A is 100 miles from Section B then a time-band of 0800 to 0859½ hours will contain different traffic unless an adjustment is made. This is simply done through the use of the journey time of the most common train type. Taking the same example, if the journey time between A and B is one hour, then the adjustment would take a time-band for Section A of 0700 to 0759½ hours as equivalent to one of 0800 to 0859½ hours for Section B.

This adjustment is particularly important for the area explanatory variables as it ensures all the geographic sections that form a particular area can be directly compared. As described in Chapter Three, the purpose of the area regression analysis was to establish whether ‘The Theory of Constraints’ was a valid approach for examining the relationship between rail capacity utilisation and performance. For this reason the time of each geographic section was adjusted from the particular primary infrastructure constraint in the sample network. The journey times of the fastest trains (generally non-stop East Coast trains) were generally used.  However on the Slow lines, the journey time of trains with the most common stopping patterns were used. The principle behind the choice of journey times to use was one of consistency between the ‘connected’ sections.

Table 5.5 gives an example of the adjusted time-bands for the sample network.

The table shows the adjusted times for the geographic sections adjacent to the primary infrastructure constraint of Welwyn Viaduct (a two-track section of line between four-track railway). This constraint is highlighted as the Woolmer Green to Welwyn Garden section[25]. Woolmer Green junction in the example is used as the base location for the 0600 to 0659½ time band. The time bands for all other locations are adjusted against this using the appropriate journey time. For example, Sandy has a 12 minute journey time to Woolmer Green Junction and Hitchin has a 4 minute journey time. The 0600 to 0659½  time-band for Sandy to Hitchin geographic section is therefore 0548 to 0647½ at Sandy and 0556 to 0655½ at Hitchin.

**Table 5.5** Example of How Time Bands are Adjusted Using the Journey Time of the Fastest East Coast Train.

| **Timing Location (Geographic Section)** | **Journey Time from Constraint** | **Adjusted Time Period** |
|---|---|---|
| Sandy (to)<br>Hitchin | 12 minutes<br>4 minutes | 0548 – 0647.5<br>0556 – 0655.5 |
| Hitchin (to)<br>Stevenage | 4 minutes<br>2 minutes | 0556 – 0655.5<br>0558 – 0557.5 |
| Stevenage (to)<br>Woolmer Green | 2 minutes<br>0 minutes | 0558 – 0557.5<br>0600 – 0659.5 |
| **Woolmer Green (to)**<br>**Welwyn Garden City** | **0 minutes**<br>**2 minutes** | **0600 – 0659.5**<br>**0602 – 0701.5** |
| Welwyn Garden City (to)<br>Potters Bar | 2 minutes<br>6 minutes | 0602 – 0701.5<br>0606 – 0705.5 |

[25] Welwyn Viaduct is in fact Woolmer Green Junction to Digswell Junction. However, in the train planning rules, activity at Digswell Junction is based on the time at Welwyn Garden City.

### 5.3.6 The Data Set for the Area Explanatory Variables

The data for the geographic sections were amalgamated to produce the data set for the area explanatory variables. As noted in the previous section, the process was greatly simplified due to the adjustment of time-bands.

The CRRD for the relevant time band for each of the sections that form the area were summed. Again this was increased by 1 so that logarithms could be used for the non-linear functional forms. The results were then divided by the total train miles to produce the dependent variable.

The CUI and HET measures for the primary constraints (LCUI and LHET respectively) were calculated using the approach outlined in Chapter Three. As explained these were simply the calculation of capacity utilisation for a small part of the sample network. This was then compared in the analysis with the CRRD per train mile for the entire area.

The calculation of the minimum ‘buffer’ for the entire area (i.e. EHET) was simply achieved by checking each of the relevant geographic sections for every train on the sample network.

### 5.3.7 The Allocation of Information to the Data Set

To summarise the methodology for producing the data set:-

- Geographic sections were identified using the method described in this Chapter i.e. they used the ‘start’ and ‘end’ points of the delay data in a ‘performance-led’ mesoscopic approach. This contrasts with Arup’s ‘infrastructure-led’ microscopic approach.
- Hourly time bands were used.
- Trains were allocated to the relevant geographic section and time-band using the described approach.
- The dependent variable was calculated by identifying the CRRD recorded against each train in each ‘cell’. This was increased by 1, to enable the use of logs. The total was then divided by the train mileage.
- The capacity utilisation explanatory variables were calculated. The formulas described in Chapter Three were applied to the identified trains.
- The ‘other’ explanatory variables were calculated using data obtained from Network Rail.

- The ‘area’ data set was produced by combining the information in the relevant geographic sections.

## 5.4 Tariff Equivalent Calculations

### 5.4.1 Overview

This section describes the methodology used to consider the implications for the pricing of congested rail networks. Once again the method adopted for the recent recalibration of the Capacity Charge provides an excellent framework. However, a number of changes have proved necessary and these are highlighted.

A key part of the analyses is the production of tariff equivalents using the values obtained from the regression work that can then be compared in detail.

### 5.4.2 Adopting the Capacity Charge Methodology

As discussed in Chapter Four, the methodology for calculating new tariff rates for the Capacity Charge is based on the Pigouvian-Knight approach to congestion pricing of charging the marginal cost of additional traffic minus the average cost (p67).

It was noted that for the exponential functional form the equation is (Faber Maunsell, 2007, p10) :-

$$\Delta D_{it} = A_i * \exp(\beta * C'_{it}) - A_i * \exp(\beta * C_{it}). \qquad (24)$$ [26]

where:

$\Delta D_{it}$ is the increase in CRRD per train mile for geographic section i and time band t. and C’ is the new CUI value following addition of one average train.

This calculation is carried out for each cell in the data set using the values for *A* and β obtained from the output of the regression analyses together with the appropriate capacity utilisation values (*C* and *C’*). For this analysis the additional CRRD was calculated using the functional form identified as the most ‘effective’.

[26] This relationship is also illustrated in Figure 4.3.

As noted previously the additional CRRD is converted into a monetary tariff using a weighted cost per delay minute for that section. It is necessary to apply a weighting where there is a mix of traffic. This is because as previously discussed (p71) different operators have  different monetary values for a minute of delay. The weighting is therefore based on the number of trains in each Service Code for the relevant section and time-band.  For example, two trains from Service Code A with a delay cost of £10 per minute and three trains from Service Code B with a delay cost of £5 per minute would lead to a weighted delay cost of £7 per minute (i.e. 10 x 2/5 + 5 x 3/5).However, due to commercial confidentiality the monetary value has not been obtained from Network Rail. Therefore, for the purposes of the analysis undertaken as part of this thesis it has had to be assumed that all delay minutes are of equal value.

Similarly, information has not been obtained for the Lateness Ratio (the ratio between delays and the Schedule 8 lateness minutes[27]) and the infrastructure fault ratio (i.e. the ratio of primary delays that Network Rail is responsible for which also includes TOC-on-TOC delay) which the tariffs are multiplied by to produce the final values. Due to these omissions the comparison between tariffs will be made using the  unadjusted values. For this reason, the results are referred to as Tariff Equivalents rather than Tariffs.

Since these issues apply equally to each capacity utilisation measure the only potential problem with this omission is where comparisons are made between the tariffs for adjacent geographic sections.

Finally, the Tariff Equivalents  have been kept at the individual value per train mile, rather than being multiplied by the number of trains to produce the total corrective tax per cell. This allows a direct comparison of the values for different geographic sections.

### 5.4.3 Calculating the Capacity Utilisation of an Additional Train

For the Traffic Intensity measure the capacity utilisation was increased by one additional ‘headway’.

[27] To recap, delay is the loss in time of a train between two timing points compared with the timetable. Lateness is the cumulative impact of delays (negatively) and allowances (positively) on a trains performance en-route when compared with the actual timetable.

For the CUI based measures the amount of time a ‘compressed’ timetable would occupy, following the addition of one  extra train[28], was calculated. This train took the form of the last train ‘on the graph’ so that there was no possible capacity benefit from ‘flighting’ for example. For the sake of consistency this was one additional ‘through’ train (i.e. no additional crossing moves were introduced to the XCUI calculations).  It was assumed that the additional train produced  leads to an increase in capacity utilisation which was equivalent to one additional ‘headway’. In other words, on a network with a 3 minute headway the addition of one  additional train would increase the CUI value in an hour by 5% (i.e. 3/60*100). This meant that the minimum capacity utilisation increase (and hence tariff) was assumed for each cell. It also ensured that there was a consistent level of increase between cells.

For the HET based measures the situation is more complicated. The inclusion of an additional train could affect the planned spacing of the existing trains significantly. This in turn will affect the calculated HET value. For example, the addition of an extra train may not be possible in any of the existing gaps and require several of them to be changed in the time period through the theoretical retiming of existing services. Alternatively, it might be possible to accommodate the new train in any or all of the existing timetable gaps. In both cases a decision has to be made about where to ‘accommodate’ the additional ‘train. This decision could have profound implications on the spacing of the new timetable and hence the final calculated tariff. There is also the distinct likelihood that different percentages could be added to the measured values for different time periods. Instead, the tariff calculation for the HET based measures clearly needs to have a single consistent approach.

The answer adopted for this thesis lies with the belief that Network Rail would logically seek to minimise reactionary delay when introducing additional traffic to the timetable. If this is the case then, according to the earlier conclusions reached in this thesis, the optimum ‘buffer’ for this new train is most likely to be one which would be obtained from even-spacing. This rationale is used to calculate the tariffs for the HET based measures.

The approach adopted is illustrated in Figure 5.1. It can be seen that the percentage value of an assumed evenly spaced ‘buffer’ for the additional train is calculated. The amount is then added to the original HET capacity

[28] For the reasons stated in this paragraph the approach adopted differs from that used for the Capacity Charge.

value for each cell in order to produce the $C'_{it}$ in Equation (24) which is then used to calculate the increase in CRRD.

**Example Four**

Assuming an existing timetable of 7 Trains (the actual spacing is unimportant).

An additional train would give an even-spacing gap for that train of 7.5 minutes (i.e. 60 / 8).

1 / 7.5 gives a reciprocal of 0.133. Dividing this by the reciprocal of the headway (assumed to be 3 minutes) gives an answer of 39.9% (i.e. 0.133 / 0.333). The additional percentage utilisation is then calculated to be 5.0% (i.e. 39.9 / 8).

**Figure 5.1** Example Four – The Derivation of the Percentage Increase for an Additional Train.

Example Four shows that for a three minute headway an additional train represents a 5% increase in capacity utilisation. It will be remembered that this is the same amount for the CUI example described earlier. This is because as demonstrated earlier in this thesis, an evenly spaced timetable produces the same CUI and HET values. In other words, one additional train equals one evenly-spaced train. In fact any number of trains for either the CUI or the HET methodologies will produce a 5% value for each additional standard train. For a four minute headway the value is 6.7%. The reason is that in each case an additional train represents the percentage value of the actual headway. The fact that the HET approach described here gives the same value as the CUI approach (and both match the value of the headway) provides reassurance that it is theoretically sound.

The next step is to divide these additional percentages by three. This takes into account the fact that in the recalibration of the Capacity Charge three-hour periods were used rather than hourly timebands. Increases of 1.67 and 2.23 % were therefore applied to the calculated capacity utilisation figures. This replicates the addition of one extra train to the three-hour periods of capacity utilisation used in the Capacity Charge recalibration (Arup, 2013).

Congestion Related Reactionary Delay was then calculated for each time band and each section using the values obtained from the regression analyses for the capacity utilisation measures. The most effective functional form was employed using the equivalent of Equation (24) to calculate the marginal increase in CRRD per train mile following the addition of a third of an extra train. In other words, if the original CUI for a geographic section with a three minute headway was 75%, this figure and 76.67% would then be applied to the relevant specification (as *C* and *C*' respectively) with the relevant values for '*A*' and '*β*'. The resulting difference in CRRD per train mile (i.e. $\Delta D_{it}$) is then used as the tariff for that geographic section and time-band[29].

### 5.4.4 The Comparison of Tariff Equivalents

The intention is to compare the calculated Tariff Equivalents for the different explanatory variables for firstly hourly time-bands and secondly for amalgamated three hour time-bands. The latter replicates the three hour time-bands used in the recalibration of the Capacity Charge. The creation of the three hourly Tariff Equivalents was achieved through a simple averaging of the relevant hourly tariffs. In other words the additional train was accounted for prior to the averaging of the hourly time periods. This approach ensures that the hourly and three hourly tariffs are consistent.

Secondly, the sectional results for any parallel Fast and Slow lines in the sample area were consolidated to produce single tariffs for each geographical section. The Tariff Equivalents for each of the two lines for a geographic section were multiplied by the proportion of train mileage that particular Fast or Slow line represented of the total, the two resulting figures were then added together to produce a new Tariff Equivalent. The use of a weighted average based on train mileage recognises that the two lines may have significantly different levels of traffic.

This consolidation was necessary for a number of related reasons:-

- Having two distinct tariffs for parallel lines might lead to the unwanted transfer of traffic from one line to the other. In other words the increased capacity advantage of having a Fast and Slow line might be lost as traffic is encouraged to concentrate on only one line.
- Having two distinct tariffs might lead to complications at the train planning stage as the Operators and Network Rail would need to take great care over the financial implications of the timetable.
- Having separate Fast and Slow line charges would substantially increase the number of tariff cells.

[29] As noted earlier in the chapter, it was not possible to apply the lateness ratio, Schedule 8 payment rate or Infrastructure Fault Ratio. For the analysis the change in CRRD per train mile is therefore used as the final tariff.

The impact of this consolidation will be returned during the discussion of the results in the next chapter.

The next step is then to further consolidate the calculated Tariff Equivalents into the equivalent of Service Code Tariff Equivalents for the sample network. This therefore achieves the same level of disaggregation reached by the Capacity Charge tariffs. Once again a weighted averaging approach is adopted to create these new Tariff Equivalents. This time the weighting is based on the mileage of each cell that forms the particular Service Code[30]. For example, the tariff for a cell that represents 1/10th of the total mileage would be multiplied by 0.1. The addition of the results for all the relevant cells together then gives the final Service Code Tariff Equivalents.

One final difference between the approach adopted for the Capacity Charge and this analysis is the issue of direction of travel. As noted in Chapter Four the Service Groups and Service Codes used in the national calibration do not distinguish between this. However, from an incentive point of view it does seem appropriate to keep the two directions separate as they will 'experience' different levels of congestion during the day. The analysis undertaken for this thesis will therefore keep the two directions of travel in the sample network separate.

The intention of the process described in this section is to allow a comprehensive comparison of different possible Tariff Equivalents for the sample network.

## 5.5 Summary

This chapter has described the methodology used to explore the relationship between capacity utilisation and performance using regression analysis. The analysis is based on the approach used during the recalibration of the Capacity Charge which provides an excellent framework.

The creation of the data sets for the analyses has then been described. This uses sample areas from one part of the British rail network. The intention is to explore the relationship between capacity utilisation and performance in

[30] As described earlier a weighted average based on train mileage was used to consolidate tariffs on parallel Fast and Slow lines. This reflects the influence of traffic volume on reactionary delay. However, for the sequential geographic sections a weighted average based on mileage reflects the importance of section length. A worked example of both approaches is given in Figure 8.1.

some detail. The transferability of the results to other congested networks can then been discussed.

Finally, the approach used to produce sample Tariff Equivalents has been explained. The intention is to examine the implications of the findings for the pricing of congested rail networks.

# Chapter 6
# The Data Set

## 6.1 Introduction

This chapter details the specific data set created to explore the relationship between capacity utilisation and performance on congested rail networks. As explained previously there are in fact two separate data sets: one designed to test the sectional and ‘other’ explanatory variables and another smaller data set designed to test the area explanatory variables. Since the smaller data set is an amalgamation of the larger one, this chapter refers to the creation of a single data set.

This chapter provides the background to the choice of the sample network and timetable used in the analysis. The key features of both are explained and the reasons behind the decisions taken are given.

## 6.2 Details of the Data Set

### 6.2.1 Overview

The decision was taken to undertake the analysis using data from just one of Britain’s rail routes. This is because, as discussed previously, the rationale is to comprehensively compare alternative explanations of the relationship between capacity utilisation and timetable performance for a single data set. The intention is then to review the findings and conclusions in order to consider their likely transferability to other rail routes. This also provides a contrast to the approach used for the calculation of the Capacity Charge in Britain, which used data for the entire British rail network but employed only one capacity utilisation measure (i.e. ‘link-only’ CUI).

Two parts of the East Coast Main Line (the primary route between Scotland and the North-East of England and London) were chosen as the basis for the data set. The East Coast Main Line (or ECML) provides an ideal choice for an investigation of the relationship between capacity utilisation and timetable performance because:-

- The route has recognised congestion issues.
- There is a variety of infrastructure which influences the utilisation of the route. This includes a number of known capacity

constraints. The applicability of the ‘Theory of Constraints’ to predicting timetable performance can thus be investigated.

- There is a significant mix of traffic types.

The use of two different parts ensures that the results are not too specific to the portion of the route chosen. These portions of the ECML were selected for their different infrastructure and traffic characteristics. They were used as the basis for comparing the different area capacity utilisation variables described in Chapter Three. The portions (or areas) were further sub-divided into geographic sections. The larger data set that was created was used for comparing the different sectional capacity utilisation and ‘other’ variables described in Chapter Three.

In all there were four areas (the two parts of the ECML were further subdivided by direction of travel) and twenty-four geographic sections. Details of the ECML and the sub-divisions used in the analysis are provided in the next section. These were produced using the methodology and rationale described in the previous chapter.

The December 2008 to May 2009 Monday to Friday (or SX) permanent working timetable was selected for the analysis. The December 2008 timetable was available and had been operated by the time that the data set for the analysis was created. This meant that the associated performance data could also be obtained. The timetable also contains a good mix of traffic, train operators and stopping-patterns for passenger trains. One reason for choosing this particular timetable, and the ECML route, is its inclusion of services for two Open Access operators (Hull Trains and Grand Central Railways). The number of passenger operators and the interaction between open access and franchised passenger services makes the timetable particularly interesting from a capacity utilisation point-of-view. There are also a substantial amount of freight paths in the ECML timetable and the issues surrounding their use are another important aspect of the analysis. Key aspects of this timetable are discussed in greater detail later in this Chapter.

### 6.2.2. The Time-bands for the Analysis

As discussed in the previous Chapter the data was divided into time-bands of one hour’s duration. However, only data for the period 0600 to 2200 hours was analysed. This gives 16 distinct time-bands. This contrasts with the eight distinct time-bands (the full twenty-four period divided into three hour periods) used in the 2013 recalibration of Britain’s Capacity Charge (Arup, 2013).

The restriction to the 0600 to 2200 time period was for a number of reasons:-

- Mid-week engineering work is common on the ECML between 2200 and 0600 hours. This is likely to increase the difference between the timetable used for the analysis and that actually operated on a day-by-day basis. This obviously has implications for the strength of the relationship between capacity utilisation and reactionary delay derived through the analysis.
- This time period is very lightly trafficked on the ECML with the majority of paths being used by ‘non-standard’[31] freight paths. This is likely to have an adverse impact on the strength of the relationships derived through the analysis. This is because the relationship between this period is likely to differ substantially from the 0600 to 2200 period when there is substantially more traffic of a more representative nature for the ECML as a whole.

The use of data for only part of the day is not considered to be an issue for this analysis. This is because the objective of understanding the relationship between capacity utilisation and reactionary delay is to consider the implications for charging for access to congested rail networks. It is therefore believed acceptable to exclude very lightly trafficked periods (which as noted have issues which potentially will affect the findings) from the analysis.

The combination of 4 areas and 16 time-bands gives a data set of 64 observations for the analysis of the area explanatory variables. The combination of 24 geographic sections and 16 time-bands gives a data set of 384 observations for the analysis of the sectional explanatory variables.

### 6.2.3 The ECML Route

Figure 6.1 shows in schematic form the southern part of the ECML, between Doncaster and London Kings Cross, within which the two areas used in the analysis are located. It demonstrates that a number of possible diversionary routes exist that avoid certain sections of the ECML. The routes via Lincoln to Newark; Lincoln and Spalding to Peterborough; via Cambridge to Hitchin

[31] The significance of ‘standard’ paths for the capacity calculations carried out for this analysis was explained in the previous chapter.

junction and the ‘Hertford Loop’ between Stevenage and London Kings Cross between them miss out a substantial part of the route.

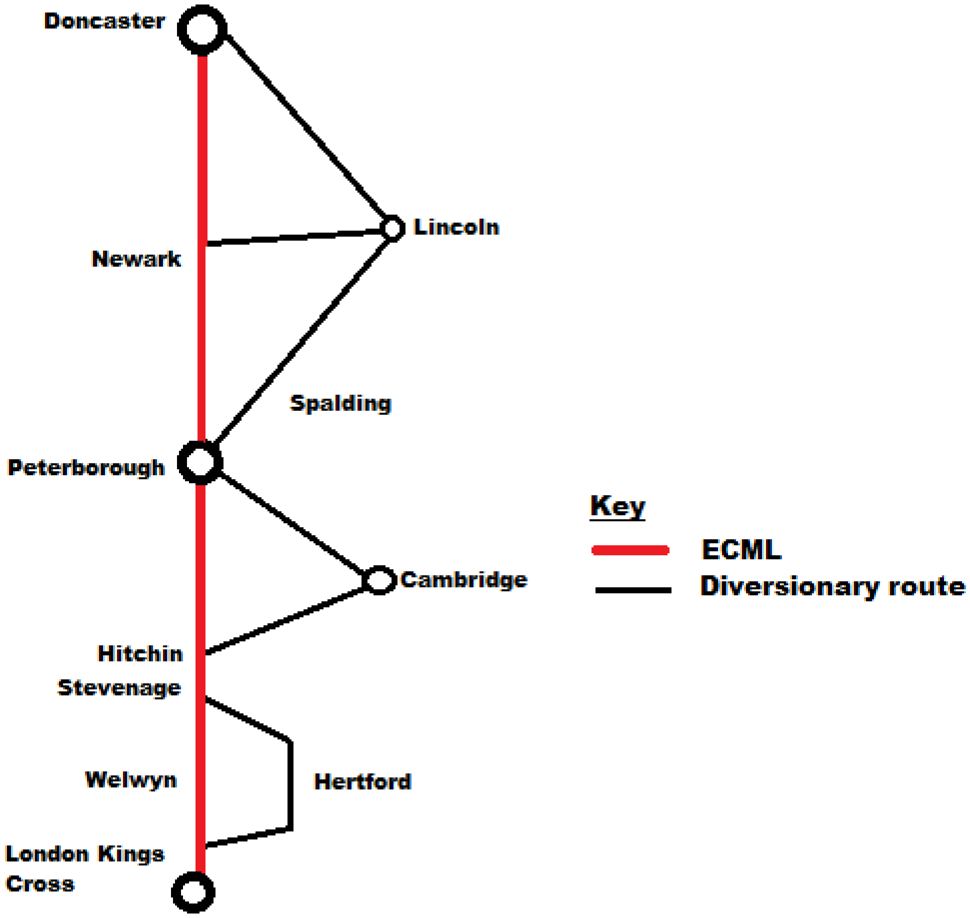


**Figure 6.1** The Southern Portion of the ECML and Associated Diversionary Routes.

However, these ‘diversionary’ routes are only suitable for certain traffic types due to their increased journey times and avoidance of many of the intermediate stations on the route. In practice this would tend to mean the transfer of freight traffic. Traffic transferring from the ECML would also have to be timed alongside existing local traffic. Reducing congestion on the ECML in this way would therefore possibly increase congestion too much on these alternate routes. This situation provides a good example of the argument between levying capacity charges on all routes and for all time periods to encourage an equilibrium in traffic levels and reducing or eliminating the charges on less congested routes to produce an increased incentive for traffic that is able to transfer. This second alternative could take the form of the de minimis approach discussed earlier in the thesis. This issue will be discussed further later in the thesis.

The Route Utilisation Strategy for the East Coast Main Line was published in February 2008 (Network Rail, 2008a) and its contents are therefore very relevant to the nature of the route at the time of the December 2008 to May 2009 timetable. Interestingly, an addendum to the ECML RUS was published in 2010 to take into account the likely impact of subsequently committed route enhancement schemes and the Intercity Express Programme on capacity issues (Network Rail, 2010a).

In his foreword to the 2008 ECML Route Utilisation Strategy Iain Coucher, the then Chief Executive of Network Rail, wrote that the ECML "is one of the busiest and most successful railway lines in Britain. As well as being an absolutely vital north-south artery for long distance traffic from London to Scotland via Yorkshire and the North East, the line serves many commuter and regional passenger markets and carries significant amounts of rail freight" (Network Rail, 2008a, p3).

The ECML broadly follows the route of the A1 and directly links the following major towns and cities with London: Edinburgh; Newcastle; Darlington; York; Leeds; Doncaster; Peterborough and Stevenage with London (Network Rail, 2008a). In addition many other parts of the country are linked due to the interconnected nature of the ECML with the rest of the British rail network.

The RUS notes that the most important use of long distance trains on the ECML is for business and leisure travel to and from London. However, there is a significant demand for travel between most key centres of population served by the route. Rail is an attractive option compared with other modes of transport and there has been strong growth in passenger travel for most long-distance flows on the ECML. For example, figures quoted in the RUS show that between 1998/99 and 2004/5 there was roughly a 40% increase in annual passenger journeys between Leeds and London (from 930,000 to 1,300,000) (Network Rail, 2008a, p29). The reason for this growth "is believed to be due to a combination of several factors, particularly economic growth and increasing road traffic congestion. On many routes the growth has been stimulated by additional services and ticketing initiatives that have been developed by operators to encourage off-peak travel" (Network Rail, 2008a, p23).

The vast majority of long distance high speed services to London are provided by the Intercity East Coast (ICEC) franchise. In addition there are the two open-access operators (Grand Central Railways and Hull Trains) which operate a number of services between Sunderland and London and Hull and London respectively. There are also long distance services on the northern portion of the ECML operated by a different franchise (Cross Country) which provides services between the South West and South East of Britain and the North West and the North East of the country.

Commuter services to London are operated by a third franchise (Thames link Great Northern[32]). These services link London with the counties of Hertfordshire, Bedfordshire and Cambridgeshire. The ECML RUS notes that there are very high levels of passenger demand in the morning and evening peak with quieter periods through the rest of the day, although Cambridge services are busy throughout the day (Network Rail 2008a). It states that for these services the market is fairly captive due to equivalent journeys by car or bus taking significantly longer. Services are divided into two distinct parts referred to as Outer Suburban and Inner Suburban. The Outer Suburban services operate between both Peterborough and Cambridge (with some services extending to Kings Lynn) and London Kings Cross. These consist of both 'semi-fast' and 'stopping' services due to the mix of station stops. The Inner Suburban services generally operate between Welwyn North or Welwyn Garden City and Moorgate station. These are 'slow' services because they generally stop at all intermediate stations.

Further north the ECML is also used by franchised regional passenger services which serve Central and Northern England and Scotland. However, traffic is heaviest on the southern portion of the ECML due to the proximity of London despite there being strong historic traffic growth on the northern part. For this reason the two areas of the ECML chosen for the analysis are on the southern portion of the route.

Parts of the route are also heavily used by freight traffic. "Approximately 30 percent of all rail freight movements in Great Britain use the ECML for at least part of their journey" (Network Rail, 2008a, p51). A huge variety of goods are transported by several different freight companies. This includes coal, steel, petroleum, container traffic, construction materials and engineering trains to support maintenance and renewal work for Network Rail. The importance of the ECML route for freight traffic reflects its strategic location in the country. For example, trains transport imported coal from various ports on the East Coast to key power stations in Yorkshire and the Trent Valley. Analysis in the RUS shows that the heaviest flow of freight trains is between York and Doncaster (as high as 30 to 40 trains a day) but

[32] Referred to in this thesis as Great Northern or GN.

there are significant flows on most of the route (for example between Doncaster and Peterborough there are from 10 to 20 trains a day and between Peterborough and London from 5 to 10 trains per day) (Network Rail, 2008a, p53). However, as noted in the previous Chapter there is a tendency for freight operators to reserve more paths in the timetable for operational flexibility than they will actually use (Network, 2008a). This provides the necessary operational flexibility they require but as discussed this introduces difficulties in the calculation of capacity utilisation figures.

The ECML is thus a strategic long-distance route which links London with Yorkshire, the North East of England and Eastern Scotland. It has a variety of significant passenger and freight flows. Most of the southern portion of the route (to the south of Grantham) has four tracks. The ‘fast’ lines allow speeds up to 125 mph and the ‘slow’ lines generally allow speeds up to 60 to 75 mph. The route north of Grantham is predominantly two-track railway but there are a number of overtaking ‘loops’ which allow faster trains to pass slower trains. The route was last modernised in the late 1980s / early 1990s (Network Rail, 2008a, p55).

The ECML route is therefore an important part of the overall British rail network to which it is highly connected. It has both high levels of traffic and a broad mixture of train types (suggesting that the number of trains and heterogeneity will both be important aspects of capacity utilisation). There are also a number of key infrastructure constraints listed by the RUS (Network Rail, 2008a). All these factors suggest that the route provides a rich subject for analysis.

There is of course the question of how representative the ECML route, and thus the findings of the analysis discussed in this thesis, is for other congested rail networks. Certainly, the aspects of the ECML described above matches those that researchers such as Abril, M., Barber, F. Ingolotti, L. Salido, M.A., Tormos, P. and Lova, A. (2008)[33] list as important factors in capacity utilisation. The ECML also has many similarities with the other main lines in Britain (e.g. the West Coast Main Line and Midland Main Line); such as a mixture of four-track and two-track railway, a mixture of long-distance high speed and shorter distance slower regional services, significant freight flows and known infrastructure constraints. It is believed therefore that although only one route has been examined in the analysis, the findings will

[33] Referred to in Chapter Three.

have general relevance for congested rail networks. The likely general relevance of the specific conclusions from the regression analysis will be discussed in Chapter Seven.

### 6.2.4 The Areas of the ECML Used in the Analysis

The selection of two parts of the ECML, rather than the whole route, ensured the analysis remained manageable. Whilst the recent recalibration of the Capacity Charge (Arup 2013) automated the calculation of CUI for the whole network, for this analysis the capacity utilisation calculations were carried out manually. This allowed the characteristics of the data set to be considered in detail, meaning the reasons behind the results obtained could be explored at some length.

The sections chosen were Loversall Carr Junction (just south of Doncaster) to Grantham Station and Sandy Station to Potters Bar Station. The principal locations for these two sections are shown in Figure 6.2. Grantham and Sandy Stations are 61.4 miles (98.8 kilometres) apart.

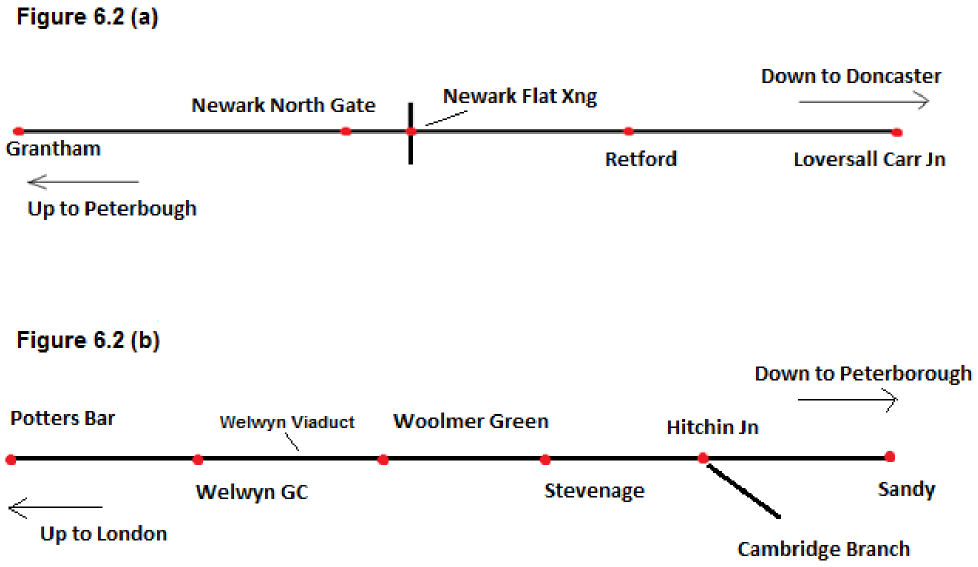


**Figure 6.2** The Two Areas of the ECML Used in the Analysis (reproduced from Haith, J., Johnson, D. and Nash, C., 2014, p27).

Loversall Carr to Grantham is a 47 mile (76 kilometre) long two-track section centred on the 'well-known' capacity constraint at Newark Flat Crossing[34], where traffic crosses the ECML between Lincolnshire and the East Midlands

[34] As illustrated in Figure 3.5.

at the same level. This part of the sample network is therefore referred to as the Newark Area (and taking into account the two directions of travel this gives the Newark Up Area and Newark Down Area). The principal locations are shown in Figure 6.2(a).

The presence of stations at Retford, Newark North Gate and Grantham produces a mix of stopping patterns for passenger services with a consequent impact on the degree of heterogeneity. The high capacity utilisation caused by an irregular stopping pattern at these stations is seen as a feature of the route. The speed differential between non-stop and stopping passenger services on this part of the route is referred to as a capacity constraint by the 2010 addendum to the ECML RUS (Network Rail, 2010a, p30). It is worth noting that  the May 2011 Timetable contained  a more regular pattern of station stops  accompanied by a significant increase in the number of train paths(Network Rail, 2010a, p8). The significant volume of freight traffic, referred to earlier, also impacts on capacity utilisation as this gives the potential for them to be ‘caught’ by the faster passenger services. This is alleviated to some extent by the presence of a number of freight overtaking ‘loops’. However, as noted at the time of the publication of the ECML RUS in 2008 these are relatively short limiting the length of freight train they can accommodate; and the entry/exit speeds are low therefore increasing the capacity utilisation by freight trains accessing these facilities. Capacity analysis for the RUS showed that although the total number of trains was fairly low in this route section a significant proportion of capacity is consumed due to the differences in speeds and calling patterns (Network Rail 2008a, p61).

Sandy to Potters Bar is a 31 mile (50 kilometre) section of mainly four-track railway. A notable exception is the case of the two-track Welwyn Viaduct (which is located between Woolmer Green Junction and Welwyn Garden City). This is a well-known infrastructure constraint that is not easy to address given the local geography. Capacity is further used by the presence of a local station (Welwyn North) on the two track section itself. The sample network is therefore referred to as the Welwyn Area (and taking into account direction of travel gives the Welwyn Up and Welwyn Down areas). The principal locations are shown in Figure 6.2(b). Not shown is the large number of ‘local’ stations served by GN’s services using the Slow lines.

In addition to the two-track Welwyn Viaduct constraint the sample area also contains Hitchin Junction. This is where Cambridge Branch traffic joins and leaves the East Coast Main Line. Until the construction of a ‘fly-over’ allowing grade-separation of conflicting moves was completed in 2013; this location was also known as a significant capacity ‘bottleneck’. The timetable chosen for the analysis therefore includes two different types of significant capacity ‘bottleneck’ in the same area. The at-grade Hitchin Junction is also of a different type to Newark Flat Crossing, since ‘local’ traffic joins (or leaves) the main line flow rather than just crossing it. The interaction between the two flows is therefore of a different nature.

Although, the level of freight traffic is not as high as in the Newark area; there is substantially more passenger traffic in the Welwyn area. This reflects the addition of commuter traffic to the long-distance passenger trains due to the closer proximity to London. Although, as noted, there is a significant mix of non-stop and stopping traffic; the presence of Fast and Slow lines means these can be separated suggesting a reduction in the level of heterogeneity caused by traffic-mix. Of course, the issue is complicated by the interaction of these two types of flow at both Welwyn Viaduct and Hitchin Junction.

One issue that does need consideration is the argument that the optimisation of the timetable can only be achieved by considering the ECML timetable as a whole. Indeed, optimising the timetables for the Loversall Carr to Grantham and the Sandy to Potters Bar sections separately is likely to produce a ‘sub-standard’ ECML timetable when it is considered overall. This is because the best use of available capacity in the two sections is unlikely to produce timings at the boundaries that match. This ‘conflict’ will potentially affect the calculated capacity utilisation and its relationship  with the level of reactionary delay. There is therefore a possible argument that two separate areas should not be used for the analysis.

However, the Capacity Charge is based on the principle of comparing the reactionary delay for discreet sections with the calculated capacity utilisation values. Any other explanatory factor in the level of delay is accounted for through the use of the fixed effects approach. This therefore suggests that the isolated area approach is appropriate. Furthermore, the inclusion of ‘other’ explanatory variables as described in Chapter Three (e.g. Average Entry Lateness and Average Distance Travelled) will it is hoped help determine the influence of the wider route on local levels of delay.

### 6.2.5 The Division into Areas and Geographic Sections

The four areas used for the analysis of the area capacity variables were those listed in the previous section (i.e. Newark Down, Newark Up, Welwyn

Up and Welwyn Down). The Welwyn areas therefore contain data for both the Fast and Slow lines.

Table 6.1 lists the twenty-four geographic sections used in the analysis of the sectional capacity variables. For the link-only capacity utilisation variables, the geographic sections exclude the station and junction node end points shown in Table 6.1. The situation is slightly more complex when junction capacity utilisation is included. The geographic boundaries for this situation are therefore clarified later in this chapter.

As referred to in Chapter Five, the recalibration of the Capacity Charge averaged utilisation for the Fast and Slow lines (Arup, 2013). Unfortunately, Arup do not give the rationale for this approach in their report. It does reduce the substantial amount of CTSs that make up the entire British rail network. The possibility of traffic ‘switching’ between Fast and Slow lines on the day of operation is also recognised. However, keeping the lines separate at this stage in the analysis for this thesis is intended to produce a more robust relationship between capacity utilisation and reactionary delay. There is also the issue that some traffic will be restricted to the Slow line due to the location of station platforms.

It will also be useful to consider any differences between calculated tariffs for parallel Fast and Slow lines and the possible reasons. If there are noticeable differences then one argument would be to keep tariffs separate since they would provide an incentive to traffic to operate on the ‘cheapest’ line. However, any difference may be actually due to the relationship between the two parallel lines (e.g. capacity utilisation on one line affects the scale of reactionary delay on the other) and therefore the tariffs should be considered in conjunction. These issues will be considered later in this thesis.

A second point worth noting is that, as described in Chapter Five, the start and end locations of the geographic sections shown in Table 6.1 all correspond with the locations used in the delay data set provided by Network Rail. The locations also correspond with mandatory timing points (i.e. locations that every train is timed at). As previously discussed this is in contrast to the approach adopted for the Capacity Charge recalibration.

**Table 6.1** Geographic Areas and Sections used in the analysis.

| Area | Section | Line |
|---|---|---|
| Newark Down | Grantham to Retford | n/a |
| | Retford to Newark | n/a |
| | Newark to Loversall Carr | n/a |
| Newark Up | Loversall Carr to Newark | n/a |
| | Newark to Retford | n/a |
| | Retford to Loversall Carr | n/a |
| Welwyn Down | Potters Bar to Welwyn | Fast line |
| | Potters Bar to Welwyn | Slow line |
| | Welwyn Viaduct | n/a |
| | Woolmer to Stevenage | Fast line |
| | Woolmer to Stevenage | Slow line |
| | Stevenage to Hitchin | Fast line |
| | Stevenage to Hitchin | Slow line |
| | Hitchin to Sandy | Fast line |
| | Hitchin to Sandy | Slow line |
| Welwyn Up | Sandy to Hitchin | Fast line |
| | Sandy to Hitchin | Slow line |
| | Hitchin to Stevenage | Fast line |
| | Hitchin to Stevenage | Slow line |
| | Stevenage to Woolmer | Fast line |
| | Stevenage to Woolmer | Slow line |
| | Welwyn Viaduct | n/a |
| | Welwyn to Potters Bar | Fast line |
| | Welwyn to Potters Bar | Slow line |

There is however three caveats with the sample network shown in Figure 6.2 and Table 6.1 and the sections used to identify the location of reactionary delay. Firstly, Newark Flat Crossing is just half-a-mile north of Newark North Gate station. Reactionary delay in the data set has either been allocated by Network Rail’s data clerks to the section between the Flat Crossing and Retford or between the Station and Retford.

The solution was to use Newark North Gate to Retford and vice versa as the geographic sections. Newark Flat Crossing is therefore effectively within a geographic section rather than at one end; though its proximity to the station makes this a moot point. The alternative would have been the creation of a very short Newark North Gate to Newark Flat Crossing section.

Secondly, Hitchin station (which is not shown in Figure 6.2 since it only has Slow line platforms) and Hitchin Junction are also very close to each other. The delay data set provided by Network Rail categorises the two as the same location. They were therefore treated as a single location (i.e. Hitchin) for the purposes of the geographic sections used in this analysis.

Finally, the Welwyn Viaduct two-track constraint is bounded by Woolmer Green Junction (shown in Figure 6.2) and Digswell Junction (not shown). It is within the Woolmer Green to Welwyn Garden City ‘performance section’ and occupies two-thirds of the length. Strictly speaking there should be short Fast and Slow line sections between Digswell Junction and Welwyn Garden City.

However, as illustrated by Table 5.4 there is a link between the timing of Fast line trains at Welwyn Garden City and Slow line trains at Digswell Junction. Creating more sections would require a number of assumptions to be made about the timing of Fast line traffic at Digswell Junction. Additionally, analysis of the delay data shows that by far the greatest influence on delay causation between Woolmer Green Junction and Welwyn Garden city is  due to the interaction between Fast and Slow line traffic. For these reasons, Woolmer Green Junction to Welwyn Garden City is represented by a single section (Welwyn Viaduct) in each direction. Due to the link between train timings described earlier, this section is in effect the Fast line between Woolmer Green and Welwyn Garden City with the addition of Slow line traffic on the viaduct itself.

The ‘linking’ of the adjacent geographic sections to take into account journey time was described in Section 5.3.5 and in particular Table 5.5. Unintentionally, it was also found that the two principal constraints in the data set (i.e. Newark Flat Crossing and Welwyn Viaduct) are almost exactly one hour's journey time apart. It will be seen that this means the results between the Newark areas and the Welwyn areas can also be compared.

## 6.3 The Sample Timetable

The issues surrounding the use of the weekday Permanent Timetable has already been discussed. The purpose of this section is to illustrate some specific features of the timetable for the two areas of the ECML chosen for the analysis. As noted in Chapter Five the December 2008 to May 2009 was the timetable chosen for the analysis.

Figure 6.3 shows the timetable graph for the Newark Down Area for the 1600 to 1700 hours time period. This illustrates the mix of traffic type on this section of the ECML and the impact on capacity utilisation. It can be seen that the graph includes four passenger trains, three freight trains (identified as such) and three crossing moves at Newark Flat Crossing (marked on the graph by 'X's). It can be seen that although there is a limited volume of trains the combination of passenger and slower moving freight trains uses a substantial amount of capacity (i.e. space on the graph).

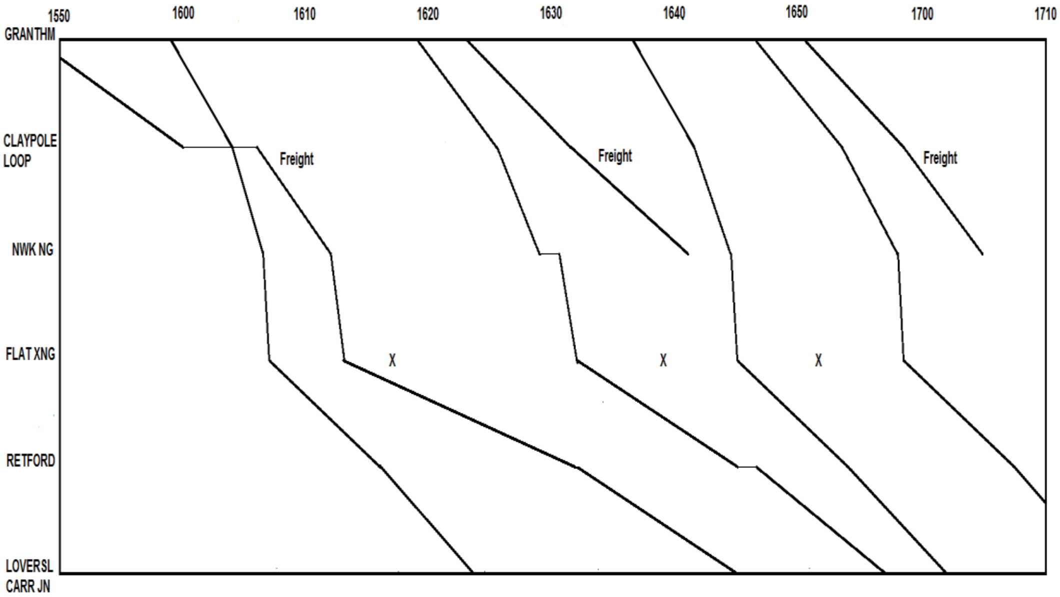


**Figure 6.3** Timetable Graph for the Newark Down Area (1600 to 1700 Time Period).

It can also be seen the first freight train, which operates the whole length of the route section needs to be 'looped' at Claypole to allow the passage of the first passenger train. In addition the first of the other freight trains which both only run for part of the route can still only just be accommodated. Although, the actual capacity utilised by the junction crossing moves are not marked (just the actual time of the crossing moves are shown), these will also clearly contribute to capacity usage on the route. The position of the first crossing move, immediately behind the planned freight train rather than in the middle of the available gap is interesting. The concepts of both vulnerable trains and different sized gaps between trains have already been discussed. Specific capacity utilisation measures to investigate this have been discussed in Chapter Three (p48). It can also be seen that the second crossing move uses the same capacity as the second freight train (i.e. both 'occupy' the gap between the second and third passenger trains). This obviously optimises capacity utilisation.

Figure 6.3 illustrates that any congestion in the Newark Area of the ECML arises principally due a mix of 'fast' and 'slow' trains rather than from the actual volume of trains.  In contrast Figure 6.4 illustrates the high volume of traffic using the Up Fast line in the Welwyn Area during the morning peak hour (0800 to 0900).

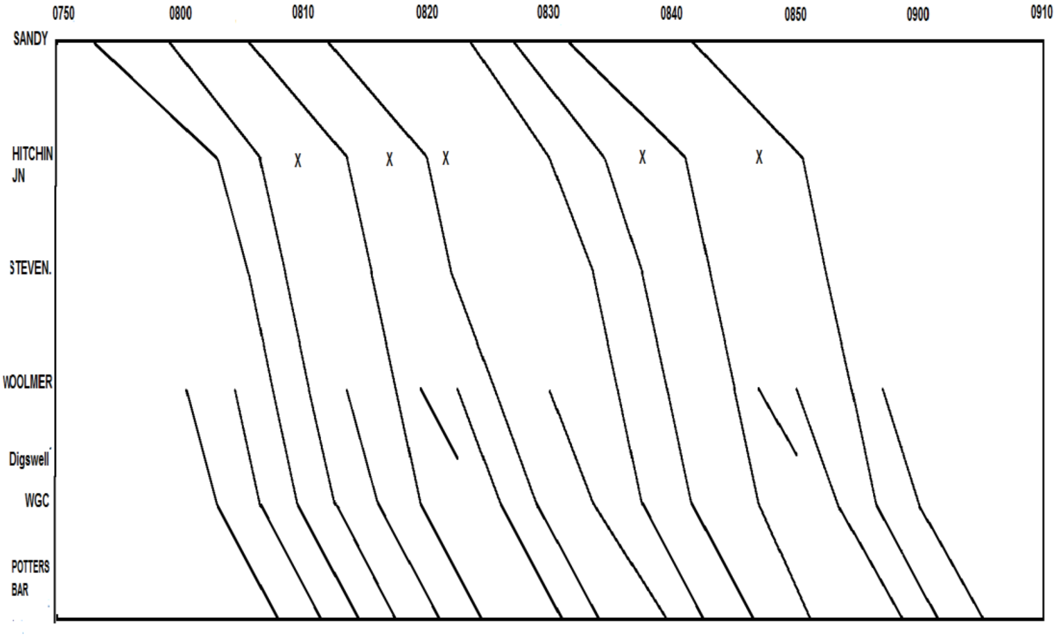


**Figure 6.4** Timetable Graph for Welwyn Up Fast Line traffic (0800 to 0900 Time Period).

As would be expected the most heavily congested part of the route shown is on Welwyn Viaduct (between Woolmer Green Junction and Digswell

Junction) where traffic on the Slow and Fast lines combine. However, many previously Slow line trains continue on the Fast line after the viaduct. The volume of capacity used prior to the viaduct and once it is reached therefore differs considerably.

However, it can be seen that the volume of traffic on the most congested part clearly influences the timetable between Sandy and Woolmer Green Junction. Although, there are considerably fewer trains there is a considerable amount of ‘bunching’. As discussed in Chapter Three, many researchers have concluded that it is the ‘buffer’ between trains that determines the level of reactionary delay. Given the fact that the HET based measures but not the CUI based measures take into account ‘buffer’ size, this could be a significant issue for the analysis.

A final point on capacity utilisation is the timing of crossing moves at Hitchin Junction which are again marked by a ‘X’. As in Figure 6.3 capacity utilisation is optimised by timing crossing moves at Hitchin Junction to use the gaps created by the Slow line traffic joining the Fast line at Welwyn Viaduct. Once again it is also noticeable that for the crossing move in the widest gap (i.e. the third crossing move), it is timed to immediately follow the passage of the first train rather than equidistant between the two trains.

Finally, Figure 6.5 shows the Up Slow line traffic in the Welwyn Area for the same morning peak period.

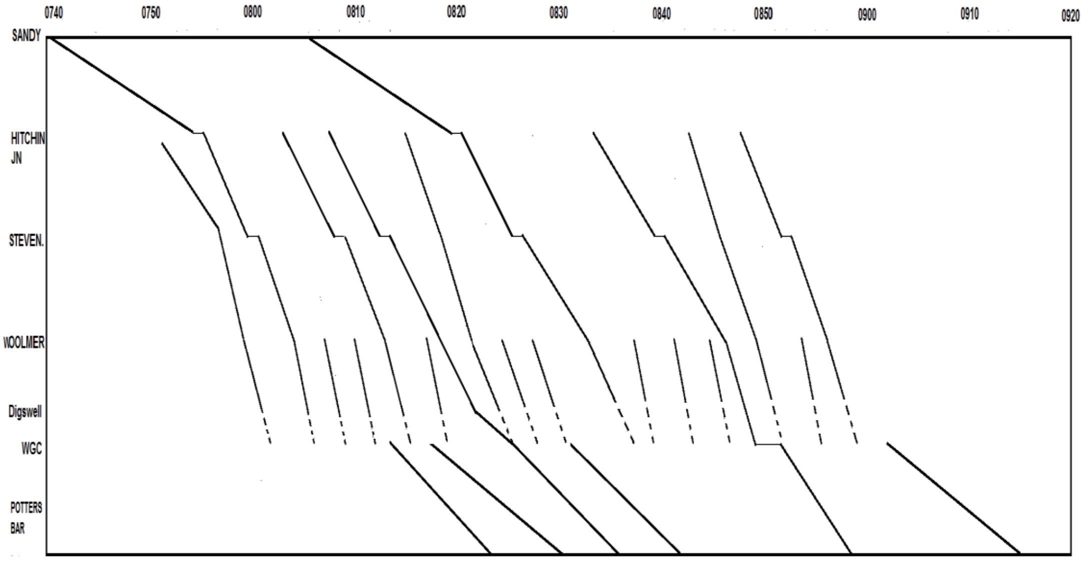


**Figure 6.5** Timetable Graph for Welwyn Up Slow line traffic (0800 to 0900 Time Period).

The lightly trafficked nature of the Up Slow line prior to the Cambridge traffic joining the ECML at Hitchin Junction is very evident. The impact of the heavily congested Welwyn Viaduct on the spacing of slow line traffic between Hitchin Junction and Woolmer Junction can also be clearly seen. For example, the first two Slow line trains are then followed over the viaduct by two Fast line trains. The impact of station stops on timetable spacing is also very obvious. For example, the fourth Slow line train on the graph stops at Stevenage but the fifth Slow line train does not. As a consequence the gap between the two trains is much smaller by the time Digswell is reached. As shown in Figure 6.4 only two of the original Slow line trains remain on this line after Welwyn Viaduct. However, at Welwyn Garden City four new trains start their journeys. Once again though, although the section between Welwyn Garden City and Potters Bar is relatively lightly trafficked there is a considerable amount of timetable ‘bunching’.

It is also important to point out a major difference between Figures 6.4 and 6.5 which is not immediately obvious given the small type face of the graph axes that has had to be used. To comfortably cover the period 0800 to 0900 at Welwyn Viaduct for Fast line traffic between Sandy and Potters Bar, a time period of 0750 to 0910 has been used. However, for Slow line traffic a time period of 0740 to 0920 has had to be used (i.e. an extra 20 minutes). The additional time used by traffic on the Welwyn Slow lines compared to the Welwyn Fast lines is therefore very clear.

Figure 6.5 therefore illustrates that capacity utilisation on the Welwyn Up Slow is effected by the mixture of stopping and non-stopping trains. However, timetable spacing is also clearly influenced by the presence of Welwyn Viaduct with its very high volume of traffic.

The three example graphs illustrate the fact that the sample data set includes congestion due to a high volume of traffic and also heterogeneity (both due to a mixture of traffic and the timetable spacing between similar trains). The sample timetables should therefore provide all the necessary elements for investigating the relationship between capacity utilisation and performance.

The need to take into account the fact that freight paths in the Permanent Timetable are not always operated has already been referred to in Chapter Five. Appendix A provides details of the freight paths that were excluded from the capacity calculations. As noted this was because they operated on 5% or less of the 110 weekdays in the December 2008 to May 2009 timetable. The statistics given in Appendix A shows that percentage operation varies widely between the different freight paths. Of the 58 freight paths listed 17 of them (or 29%) were excluded from the data set.

## 6.4 The Reactionary Delay Data Set

### 6.4.1 Overview

This next section discusses the features of the ECML CRRD data which was used for the analyses described in this thesis.

The ‘performance-led’ approach for this analysis compared to the ‘infrastructure-led’ approach for the recalibration of the Capacity Charge in Britain was discussed in the previous Chapter. The mesoscopic approach compared to a microscopic approach was also outlined. It was noted that there are some disadvantages but it is firmly believed that these are out-weighed by the advantages of this approach. A ‘performance-led’ mesoscopic approach also provides an alternative perspective to the recalibration of the Capacity Charge.

The next section will discuss the key features of the CRRD data. In considering the data, it should be remembered that in Chapter Five it was noted that it was necessary to partially allocate CRRD for Hertford Loop traffic on the ECML in the Stevenage – Woolmer Green Junction on a pro-rata basis. This led to 10% of the affected delay being allocated onto the ECML. Only a very small amount of delay was affected in this way. However, this explains why the delay minutes for the Welwyn Slow lines are shown as ‘odd’ fractions of whole minutes.

### 6.4.2 High Level Analysis of the Delay Data

Table 6.2 shows the total CRRD included in the data set for the analysis. The reactionary delay recorded for Welwyn Viaduct itself is included in the Fast line totals. The fractions in the totals for the Slow lines reflect the ‘Langley Junction’ issue referred to in the previous section. As might be expected, the reactionary delay recorded in the Welwyn Area overall exceeds that for the Newark area. Although, the former is a shorter section of the ECML it does have a greater volume of traffic due to its closer proximity to London. It can be seen that both the Welwyn Fast lines generate more reactionary delay than the Welwyn Slow lines. As noted previously more traffic uses the Fast lines and as noted these totals include the delay for the Welwyn Viaduct constraint.

**Table 6.2** Minutes CRRD Used in the Analysis (Analysis by Author)

| Area | Reactionary Delay (minutes) | % of Total |
|---|---|---|
| Newark Down | 2973 | 16.3 |
| Newark Up | 3535 | 19.3 |
| Welwyn Down Fast | 3261 | 17.8 |
| Welwyn Down Slow | 1949.7 | 10.7 |
| Welwyn Up Fast | 4087 | 22.3 |
| Welwyn Up Slow | 2483.8 | 13.5 |
| Overall | 18289.5 | 100.0 |

One final observation is that in each case, the reactionary delay is greater for the Up direction than the respective Down direction. Although, it could be surmised that this is the influence on timetable performance of traffic heading towards the southern terminal of the line (i.e. London Kings Cross); this still applies for the Newark area which is a much greater distance from London than the Welwyn area. Another possibility is that the greater level of reactionary delay on the Up is due to the greater distance travelled by trains than in the Down direction. As discussed in Chapter Three, 'Average Distance Travelled' is one of the variables used in this analysis. This possibility will therefore be investigated.

Table 6.3 provides details on the amount of reactionary delay per record using the mean, median, mode and standard deviation. It will be seen, as indicated by a '*' that the minutes reactionary delay for the Welwyn Slow Lines and therefore the overall total do not match the totals given in Table 6.2. This is because the 'Langley Junction' CRRD has been excluded from the analysis, so that only 'pure' ECML reactionary delay is included in the break-down.

Analysis of the table shows that once the number of reactionary delay minutes is divided by the number of records, the picture given by Table 6.2 changes somewhat. The mode for each line (and overall) is the same at 3 minutes of reactionary delay per incident. The median for the Welwyn lines (and overall) is also 3 minutes of CRRD per record. However, it can be seen that the median is higher for the Newark lines at 4 minutes. The greater number of records for the two Welwyn Fast lines also means that their mean level of reactionary delay is in-fact lower than that seen on the Newark lines. Overall, whilst the total reactionary delay is higher in the Up direction (as discussed with reference to Table 6.2) it will be seen that this is due to the greater number of records of reactionary delay. The mean minutes of reactionary delay are also relatively low and it can be seen from the standard deviation that for those with the highest means (both Newark lines and the Welwyn Down Slow line) this is due to a number of records with relatively high levels of reactionary delay.

**Table 6.3** Analysis of the Size of Reactionary Delay per Record (Analysis by Author).

| Area | CRRD minutes | Number of Records | Mean Minutes per record | Median Minutes per record | Mode Minutes per record | Standard Deviation |
|---|---|---|---|---|---|---|
| Newark Down | 2973 | 591 | 5.0 | 4 | 3 | 4.0 |
| Newark Up | 3535 | 730 | 4.8 | 4 | 3 | 3.7 |
| Welwyn Dn. Fast | 3261 | 828 | 3.9 | 3 | 3 | 2.7 |
| Welwyn Dn. Slow | 1911* | 469* | 4.1 | 3 | 3 | 5.3 |
| Welwyn Up Fast | 4087 | 1091 | 3.7 | 3 | 3 | 2.4 |
| Welwyn Up Slow | 2480* | 673* | 3.7 | 3 | 3 | 1.7 |
| Overall | 18247* | 4382* | 4.2 | 3 | 3 | 2.9 |

In summary, the overall mean, median, mode and standard deviation show that reactionary delay per record per train is relatively low. It will be seen that this finding has important consequences for the interpretation of the results and the subsequent conclusions contained later in this thesis.

## 6.5 Creation of the ECML Variables

### 6.5.1 Overview

The methodology used to create the dependent and explanatory variables for the regression analyses was described in Chapter Five. The purpose of this section is to provide information specific to the ECML sample network.

A necessary first step in the calculation of the dependent and explanatory variables was the identification of the trains 'present' in each cell in the sectional data base. This was carried out using the ECML timetable data supplied by Network Rail. The methodology for allocating trains that 'straddled' time-bands or crossed between Fast and Slow lines was discussed in the previous Chapter.

### 6.5.2 The Dependent Variable

For each cell in the sectional data set the CRRD was calculated using the collated information on the trains present and the delay data. The total was then, as previously discussed, increased by one, due to the logarithmic nature of some of the functional forms being examined. As noted this was the solution adopted for the recalibration of the Capacity Charge and is a standard solution to the problem (Arup, 2013). The total train mileage for each cell was calculated through reference to the ECML track diagrams (Trackmaps, 2005). The CRRD+1 was then divided by the train mileage to produce 'CRRD+1 per train mile'. This Dependent variable is referred to throughout this thesis as RD1TM.

The dependent variable for the area data set was simply created by combining the relevant delays and mileages from the sectional data set. As noted earlier, the area data set consists of 64 individual cells (compared to the 384 in the sectional data set); and consists of the Newark Down, the Newark Up, the Welwyn Down and the Welwyn Up areas. In the case of the latter, the Fast and Slow line data were combined to give overall CRRD per Train Mile in the Welwyn area by direction.

### 6.5.3 The Explanatory Variables

The explanatory variables were calculated using the formulae discussed in Chapter Three and the methodology outlined in Chapter Five.

As a reminder the capacity utilisation variables are shown in Table 6.4.

The Traffic Intensity variable was the easiest of the three types of capacity utilisation variables to calculate. The numbers of trains in each time period and geographic section were ‘compared’ with the theoretical maximum number of trains on the section based on the timetable planning headway. This was three minutes for the majority of the Welwyn Fast line sections and four minutes for the Welwyn Slow line sections and Newark area sections.

The calculation of the two CUI based capacity utilisation variables (OCUI and XCUI) was more complicated. The ‘compression’ process for both variables was carried out again using the relevant train path details and the appropriate ECML timetable planning rules (Network Rail, 2009b). As discussed in Chapter Three, the latter determines the minimum gap between successive trains in the ‘compressed’ timetable.

**Table 6.4** Reminder of the Sectional Capacity Utilisation Variables used in the Analysis

| Abbreviation | Measure | Equation (in Chapter 3) |
|---|---|---|
| I | Traffic Intensity | (2) |
| OCUI | Link CUI | (3) |
| XCUI | Junction & Link CUI | (4) |
| OHET | Link HET | (8) |
| AHET | Arrival HET | (9) |
| XHET | Junction & Link HET | (10) |
| VHETB | ‘Vulnerable’ HET (before) | (11) |
| VHETF | ‘Vulnerable’ HET (following) | (12) |

In the case of links with an adjacent junction the XCUI value was calculated using the appropriate junction margins. The links with associated junctions are highlighted later in this section. For consistency the same ones were obviously used by the XHET variable as well.

The HET based capacity utilisation variables were calculated using the approach described in Chapter Three. In their specific case the calculation was undertaken by creating spreadsheets with the appropriate planned times noted. The gaps between these times were then calculated. Once again the relevant planning headways and junction margins were obtained from the Timetable Planning Rules (Network Rail, 2009b). The ‘adjustment’ process to ensure that a consistent denominator was used is as described in Chapter Three.

The calculation of the Junction and Link capacity utilisation measures (XCUI and XHET) require special mention. The first step was to identify which geographic sections required the calculation of XCUI and XHET figures due to the proximity of Newark Flat Crossing, Hitchin Junction and the junctions either end of the Welwyn Viaduct constraint (as noted previously any other switches and crossings in the sample network are treated as special cases using the appropriate timetable planning rules). Table 6.5 shows the geographic sections for which XCUI and XHET figures were calculated.

**Table 6.5** Geographic sections for which XCUI and XHET were also calculated.

| **Geographic Section** | **Line** | **Constraint accounted for** | **Position in Link** |
|---|---|---|---|
| Newark - Retford | Down | Newark Flat Crossing | Within |
| Retford - Newark | Up | Newark Flat Crossing | Within |
| Potters Bar – Welwyn Garden City | Down Fast | Welwyn Viaduct | End |
| Potters Bar – Welwyn Garden City | Down Slow | Welwyn Viaduct | End |
| Stevenage – Woolmer Green Jn | Up Fast | Welwyn Viaduct | End |
| Stevenage – Woolmer Green Jn | Up Slow | Welwyn Viaduct | End |
| Hitchin – Sandy | Down Slow | Hitchin Junction | Start |
| Stevenage – Hitchin | Down Fast | Hitchin Junction | End |
| Sandy – Hitchin | Up Fast | Hitchin Junction | End |
| Sandy – Hitchin | Up Slow | Hitchin Junction | End |

It can be seen that in total ten of the twenty-four geographic sections have had the impact of junction moves on capacity utilisation included. The basic principle has been to take into account the impact of a junction on the link approaching the constraint. This is based on the rationale that ‘approaching’ traffic would be most likely to incur reactionary delay due to congestion ahead of it.

However, as can be seen in Table 6.5 this rule has not been universally applied to the ten geographic sections. As previously discussed, the proximity between Newark North Gate station and Newark Flat Crossing means that the most appropriate sections to account for the latter is within the Retford to Newark North Gate ones. The issue of Welwyn Viaduct and Digswell Junction  has already been discussed. It can be seen that the capacity utilisation of Digswell Junction (which forms one end of the Viaduct) is accounted for at the end of the Potters Bar to Welwyn Garden City sections. This again reflects how the operation of this junction is accounted for in the train planning rules (Network Rail, 2009b).

It does mean however that the approach taken assumes that reactionary delay caused due to approaching congestion on Welwyn Viaduct occurs between Potters Bar and Welwyn Garden City rather than up to Digswell Junction. However, given that the Welwyn Viaduct sections themselves include all reactionary delay listed as occurring between Woolmer Green Junction and Welwyn Garden City, this is considered acceptable.

Finally, it can be seen that three of the four main lines at Hitchin Junction follow the rule of accounting for the impact of the junction in the ‘approach’ geographic sections. However, this is not the case for the Down Slow. The reason for including Hitchin Junction in the Hitchin to Sandy section is due to the layout of the junction and the fact that Hitchin Station and Hitchin Junction are treated in the timetable and delay data as the same point. In the Down direction, Hitchin Station is located immediately in advance of the crossover to the Cambridge branch at the junction. Departing the station, main line traffic stays on the Down Slow section to Sandy whilst branch line traffic crosses Hitchin Junction. It therefore follows that any delay to Down Slow traffic due to the operation of the junction is most likely to occur in the Hitchin to Sandy section.

The ‘other’ capacity utilisation variables (i.e. Time Before, Section Before and Section Following) were calculated using  the approach adopted for the relevant capacity utilisation variable.

The ‘other’ explanatory variables described in Chapter Three were also calculated for each cell in the sectional data set. The data for the ECML (and rest of the network as appropriate) was used as described in Table 6.6. It can be seen that four of the five ‘other’ explanatory variables use individual train timings. These records show the entirety of each train path. The fifth variable (‘Average Entry Lateness’) used the lateness data for each timing point on the ECML sample networks.

**Table 6.6** Data Sources for the ‘Other’ Explanatory Variables used in the Sectional Data Analysis.

| Variable | Data Source(s) |
|---|---|
| Timetable Complexity<br>(TTC) | Individual TrainTimings<br>(Network Rail Timetable Data) |
| Average Entry Lateness<br>(AEL) | Individual Lateness Records<br>(Network Rail ECML Performance Data) |
| Stability<br>(STAB) | Individual Train Timings<br>(Network Rail Timetable Date) |
| Average Transit Time Variation<br>(ATV) | Individual Train Timings<br>(Network Rail Timetable Data) |
| Average Distance Travelled<br>(ADT) | Individual Train Timings<br>(Network Rail Timetable Data / Railway Track Diagrams) |

The three area capacity utilisation variables (i.e. LCUI, LHET and EHET) were also calculated using the methodology described previously. In the case of LCUI and LHET (or Local CUI and Local HET respectively) the relevant calculations were carried out for the identified primary area infrastructure constraints (i.e. the Flat Crossing for the Newark area and Welwyn Viaduct for the Welwyn Area). With Newark Flat Crossing being a node rather than a link it is necessary to provide some further explanation on its calculation. The capacity utilisation was calculated for the actual location rather than for the ‘approach’ links. For LCUI, this meant the relevant headways and margins were used to ‘compress’ the timetable for Newark Flat Crossing itself. For LHET, this meant that the ‘gaps’ in the timetable at Newark Flat Crossing itself were calculated. However, the XHET approach was still used (i.e. where a crossing move was timetabled before the train in question, the gap to the previous ‘through’ train was also counted). This maintained consistency with the approach adopted for modelling capacity utilisation at a sectional level. For EHET (or Expanded HET) the minimum gap to a preceding train, wherever that occurred in the area in question, was used.

A sample of the information contained in the data sets is provided in Appendix B. This shows an example of the calculated dependent variable and the capacity utilisation independent variables for the sectional and area data sets and the ‘other’ independent variables for the sectional data set.

## 6.6 Summary

This chapter has described the two sample parts of the East Coast Main Line used to create the data sets for the analysis of the relationship between capacity utilisation and reactionary delay. The rationale behind the choice of the two parts of the ECML has been discussed in some detail. The different characteristics of the two sections have been explained. In particular the fact has been pointed out that whilst the Newark area has mixed traffic but a relatively low number of trains, the Welwyn area has less of a mix of traffic but a high volume of trains.

The Congestion Related Reactionary Delay Data for the ECML used as the dependent variable has been described. The low number of delay minutes per observation has been highlighted. It will be seen that this has a significant influence on the conclusions of the analysis.

The results of the regression analyses will be detailed in the next chapter. The immediate conclusions that can be drawn from these results will be discussed and this will be illustrated by examples drawn from the data set. The conclusions that can be drawn about the relationship between capacity utilisation and reactionary delay will then be outlined. An important part of this discussion will be the potential transferability of any findings for the sample network to congested rail networks as a whole. Chapter Eight will then use these conclusions as a basis for considering the most appropriate means of charging for access to congested rail networks.

# Chapter Seven
# Results

## 7.1 Introduction

This chapter discusses the outcome of the regression analysis of the data set described in Chapter Six of this thesis. The variables described in Chapter Three are compared. These were summarised in Tables 3.2 and 3.5. Reference to the Abbreviations section of this thesis provides the full name of each of the variables concerned. As described in Chapter Five, the methodology broadly follows the same approach adopted for the 2013 recalibration of the Capacity Charge.

The reasons for the results are then discussed, illustrated by some examples from the data set. Finally, the implications of these findings are outlined and how representative they are of congested networks as a whole is considered.

## 7.2 Regression Results for the Sectional Capacity Measures

### 7.2.1 Identification of the most appropriate functional form

As discussed in Chapter Five, the use of linear and non-linear functional forms requires a transformation of the data set to allow a direct comparison. A Box-Cox transformation (as described by Dougherty, 2011) was the method adopted. The procedure compares the 'fits' of the linear and logarithmic specifications using the residual sums of the squares.[35]

Regressions were carried out for the four different options (i.e. one-way fixed effects; two-way fixed effects; one-way random effects and two-way random effects[36]) for the different functional forms and the different capacity utilisation explanatory variables.

The resulting residual sums of the squares were then compared. For every explanatory variable and every option examined, the functional forms with a

[35] See Section 5.2.3 for a detailed explanation of the process.

[36] See Sections 5.2.4 and 5.2.5 for a detailed explanation of these alternative approaches.

logarithmic dependent variable were found to have substantially lower residual sums of the squares than those with linear dependent variables.

Table 7.1 shows (for reasons of brevity) the residual sums of squares for just the one-way fixed effects results for each of the five functional forms being considered[37]. The difference in the level of the residual sums of squares can easily be seen. The results for the other options give a similar result. It can therefore be concluded that the non-linear functional forms are more appropriate than the linear ones. The Exponential and 2nd Order Approximation (logarithmic) functional forms were therefore taken forward for further consideration.

**Table 7.1** Comparison of the Residual Sums of Squares for the Five Functional Forms (One-Way / Fixed Effects).

| **Capacity Variable** | **Linear (linear)** | **Quadratic (linear)** | **2nd Order Approx. (linear)** | **Exponential (logarithmic)** | **2nd Order Approx. (logarithmic)** |
|---|---|---|---|---|---|
| Intensity | 601.89 | 599.26 | 597.93 | 323.99 | - |
| OCUI | 598.79 | 598.43 | 598.42 | 323.39 | 324.94 |
| XCUI | 595.20 | 593.39 | 592.88 | 322.06 | 319.08 |
| OHET | 568.19 | 572.21 | 567.96 | 286.44 | 289.58 |
| AHET | 583.86 | 589.74 | 581.30 | 302.10 | 299.51 |
| XHET | 535.05 | 562.16 | 556.75 | 283.49 | 277.52 |
| VHETB | 566.40 | 570.10 | 566.26 | 288.99 | 281.64 |
| VHETF | 551.58 | 555.30 | 551.03 | 282.12 | 293.07 |

(note a result for Intensity using the 2nd Order Approx. form cannot be calculated due to perfect colinearity).

### 7.2.2 Decision Between ‘Fixed’ and ‘Random’ Effects

It is then necessary to make a decision between fixed and random effects for each of the two (logarithmic) functional forms. As discussed in Chapter 5, the standard approach is to use a Hausman Test. The analysis was carried

[37] The other Residual Sums of Squares results are shown in Appendix B.

out using non Box-Cox transformed data. This is necessary at some point because, as pointed out by Dougherty (2011), using the transformations means that the size of the actual coefficients will differ. The results in the form of Chi Square values are shown in Table 7.2.

**Table 7.2** Chi Square Statistics Calculated by the Hausman Test for the Sectional Variables.

| Capacity Utilisation Variable | One Way Exponential | Two Way Exponential | One Way 2nd Order Approx (Logarithmic) | Two Way 2nd Order Approx (Logarithmic) |
|---|---|---|---|---|
| Intensity | 0.260 | 5.069 | - | - |
| OCUI | 0.035 | 6.592 | 0.337 | 12.479 |
| XCUI | 1.411 | 6.489 | 3.591 | 13.717 |
| OHET | 1.812 | 1.111 | 3.055 | 0.810 |
| AHET | 1.895 | 1.485 | 2.881 | 1.146 |
| XHET | 0.441 | 4.249 | 0.752 | 2.778 |
| VHETB | 1.310 | 2.173 | 2.574 | 1.293 |
| VHETF | 1.385 | 1.802 | 2.544 | 1.079 |

(note a result for Intensity using the 2nd Order Approx. form cannot be calculated due to perfect colinearity).

The critical value for the Exponential functional form is 3.815 at a 95% Confidence Interval. The critical value for the Second Order Approximation (Logarithmic) form is 5.992. It can therefore be seen that in the majority of cases the null hypothesis cannot be rejected. Dougherty (2012) explains that the null hypothesis is that both random effects and fixed effects are consistent. However, the use of fixed effects would be inefficient compared to random effects as time-invariant variables cannot be used and a number of degrees of freedom are lost due to the additional requirement to calculate dummy variables. The recommendation in the case of the null hypothesis not being rejected is that random effects is adopted.

Although, it can be seen that the vast majority of the capacity variables and different approaches favour random effects under the Hausman Test, the decision has however been taken to adopt a fixed effects approach. This is for the following reasons:-

- The null hypothesis is that both random effects and fixed effects are appropriate but the former is more efficient. Given the fact that the Capacity Charge carried out at a national level ultimately selected a Fixed effects approach, it is appropriate to retain consistency.
- Following on from the first point, the likelihood that there is a relationship between unobserved effects and capacity utilisation is highly intuitive. This is because no matter what the local circumstances, reactionary delay can only be triggered by the presence of traffic.
- As noted on page 99 of this thesis, Dougherty (2011, p525) states that if the data set is non-random then fixed effects should be used. As described earlier, the two sample ECML areas although carefully chosen cannot necessarily be considered random. They were specifically chosen to test the ability of various capacity utilisation measures to predict reactionary delay on areas with known congestion issues.
- In the cases where fixed effects has been identified, i.e. where the null hypothesis is rejected, the use of random effects will produce biased results.

It is strongly believed for these reasons that the risk of inefficiency from adopting fixed effects is therefore acceptable.

### 7.2.3 Decision Between ‘One-Way’ and ‘Two-Way’ Models

Table 7.3 shows the t-statistic results for the one and two-way approaches for the Exponential functional form.  Table 7.4 shows the F-test of Joint Significance results for the Second Order Approximation (logarithmic) form. Table 7.5 shows the adjusted R-squared results for both functional forms. In line with the decision taken in Section 7.2.2 all results are for fixed effects.

The critical t-value for a 95% confidence limit for this size of data set is 1.966. Table 7.3 therefore shows that for the Exponential functional form all capacity utilisation variables considered for both the  one-way and two-way approaches are significant. It is necessary to note though that the size of the t-statistic cannot be used to choose between the two approaches. This is because fewer coefficients are being estimated by the one-way models.

**Table 7.3** Comparison of the Sectional t-statistic Results for the One-Way and Two-Way approaches (Exponential Functional Form)

| Capacity Utilisation Variable | Exponential One-Way Fixed Effects | Exponential Two-Way Fixed Effects |
|---|---|---|
| Intensity | 5.054 | 2.288 |
| OCUI | 5.124 | 2.982 |
| XCUI | 5.277 | 3.140 |
| OHET | 8.714 | 6.295 |
| AHET | 7.308 | 4.837 |
| XHET | 8.971 | 6.732 |
| VHETB | 8.491 | 6.104 |
| VHETF | 8.190 | 5.891 |

**Table 7.4** Comparison of the Sectional F-Test of Joint Significance Results for the One-Way and Two-Way approaches (Second Order Approximation (Logarithmic) Functional Form).

| Capacity Utilisation Measure | One-Way Fixed Effects | Two-Way Fixed Effects |
|---|---|---|
| Intensity | - | - |
| OCUI | 12.986 | 5.742 |
| XCUI | 16.736 | 7.496 |
| OHET | 37.900 | 19.757 |
| AHET | 30.309 | 13.979 |
| XHET | 47.846 | 28.670 |
| VHETB | 37.203 | 19.621 |
| VHETF | 35.173 | 18.349 |

(note a result for Intensity using the 2nd Order Approx. form cannot be calculated due to perfect colinearity).

The critical F-test of Joint Significance value for a 95% Confidence Limit is 3.019. Table 7.4 therefore shows that once again all capacity utilisation measures for both the one-way and two-way approaches have significance with this functional form.

Table 7.5 shows that although all the capacity utilisation variables have reasonable adjusted R-squared results there is some difference between them. This table will be used to inform the choice about the most effective capacity utilisation measure. Both the one-way and the two-way approaches have reasonable adjusted R-squared values.  It can be seen that the one-way approach tends to produce better results for the HET based capacity utilisation measures, whilst the two-way approach is better for Intensity and the CUI measures. However, as discussed by Arup (2013, pp23-24) a two-way approach (i.e. where individual coefficients are given for time as well as geographical variance) does not appear logical. Arup (2013, p21) also note that the two-way models for the preferred functional forms “imply a u-shaped or downward sloping curve which seem counter-intuitive”. Therefore for the purposes of this analysis the one-way approach is preferred.

**Table 7.5** Comparison of the Sectional Adjusted R-squared Results for the One-Way and Two-Way approaches.

| **Capacity Utilisation Variable** | **Exponential One-Way** | **Exponential Two-Way** | **2nd Order Approx. (logarithmic) One-Way** | **2nd Order Approx. (logarithmic) Two-Way** |
|---|---|---|---|---|
| Intensity | 0.432 | 0.440 | - | - |
| OCUI | 0.433 | 0.446 | 0.429 | 0.447 |
| XCUI | 0.435 | 0.448 | 0.439 | 0.452 |
| OHET | 0.498 | 0.490 | 0.491 | 0.484 |
| AHET | 0.470 | 0.468 | 0.473 | 0.469 |
| XHET | 0.503 | 0.498 | 0.512 | 0.504 |
| VHETB | 0.493 | 0.487 | 0.489 | 0.483 |
| VHETF | 0.487 | 0.484 | 0.485 | 0.480 |

(note a result for Intensity using the 2nd Order Approx. form cannot be calculated due to perfect colinearity).

### 7.2.4 Autocorrelation and Heteroskedasticity

The use of panel data in the Eviews software does not currently fully support tests for autocorrelation and heteroskedasticity.

Autocorrelation was tested for using a Durbin-Watson test and a Lagrange-Multiplier test. There was no evidence of autocorrelation. However, since these standard tests for autocorrelation assume that that data is 'stacked' continuously i.e. the observation at the end of one sub-division is immediately followed by the observation in the next sub-division there is a potential problem with using these tests with panel data. Therefore, although not ideal, visual inspection of the residuals was also used. This approach also produced no evidence of auto-correlation.

For heteroskedasticity dummy variables were used to reflect changes in geography (i.e. the equivalent of a fixed effect one-way model) in a new data set (i.e. non-panel data) in order to permit the use of appropriate Heteroskedasticity tests. Once again due to the nature of panel data it is not clear whether this approach is appropriate. It is therefore recognised that it is not ideal. However, it will be seen that the conclusions drawn from the analysis are made on the basis of both homoskedastic and heteroskedastic conditions.

Evidence of heteroskedasticity was investigated using a White test. Bearing in mind the caveat given above, heteroskedasticity was identified for each of the relationships investigated. This is contrary to the findings of the Capacity Charge recalibration which instead found that the data was homoskedastic (Arup 2013, p25). It is therefore worth considering this apparent anomaly. There is some logic to the relationship between capacity utilisation and reactionary delay exhibiting heteroskedasticity. As discussed previously in this thesis, reactionary delay is in the first instance triggered by a primary incident which can occur at any time; however at higher levels of congestion there is more traffic to incur and further propagate reactionary delay. This therefore suggests a greater variability in the level of reactionary delay at higher levels of capacity utilisation i.e. the relationship is heteroskedastic. The homoskedasticity of Arup's analysis may reflect the wider time bands used in their analysis which will have reduced the overall variation in the data.

Heteroskedasticity was accounted for in the regression analysis undertaken for this Thesis using a White Heteroskedastictiy Consistent Covariance Matrix Estimator approach. This is suitable for instances when the heteroskedasticity is of unknown form (QMS, 2010, p33). This adjustment affects the level of the t-statistic and these revised figures are reproduced in Table 7.6. It will be seen that both one-way and two-way models have been revised despite the test only being carried out for the former. This takes into account the possibility that the two-way approach also suffers from heteroskedasticity and means that the results can be compared.

Table 7.6 demonstrates that following adjustment to account for the possibility of heteroskedasticity, all the capacity utilisation measures remain significant.

**Table 7.6** Comparison of the Sectional t-statistic Results (Adjusted for Heteroskedasticity) for the One and Two-Way approaches (Exponential Functional Form).

| **Capacity Utilisation Variable** | **Exponential One-Way (White Adjustment)** | **Exponential Two-Way (White Adjustment)** |
|---|---|---|
| Intensity | 4.314 | 2.234 |
| OCUI | 4.378 | 2.795 |
| XCUI | 4.597 | 2.938 |
| OHET | 7.931 | 6.340 |
| AHET | 6.344 | 4.716 |
| XHET | 8.601 | 6.432 |
| VHETB | 7.574 | 6.090 |
| VHETF | 7.392 | 6.064 |

### 7.2.5 Choice Between Functional Forms and Capacity Variables

Table 7.7 repeats the  adjusted R-squared results for the fixed effects one-way approach model. It can be seen that there is an important relationship between each capacity variable and reactionary delay. The data in Table 7.7 however also gives the opportunity to decide which is the most appropriate functional form and the most effective capacity utilisation explanatory variable for the sectional data set.

The results show that the two functional forms both have higher adjusted R-squared scores for four of the eight capacity utilisation measures. It can be seen though that the Second Order Approximation (logarithmic) functional form does produce the highest overall adjusted R-squared result (0.512 for the XHET capacity utilisation measure). Based on the adjusted R-square results alone the Second Order Approximation (logarithmic) functional form can be said to 'best' describe the relationship between capacity utilisation and reactionary delay. It can be seen though that the difference between the results for the two functional forms is relatively small. Table 7.7 also shows that Intensity has the lowest adjusted R-squared value. This is followed by OCUI and then XCUI. This follows expectations as XCUI includes the impact of junction moves, rather than being solely link-only based.

**Table 7.7** Adjusted R-squared Sectional Results for the Fixed Effects One-Way Models.

| **Capacity Utilisation Variable** | **Exponential Adjusted R-squared** | **2$^{nd}$ Order Approx. (Logarithmic) Adjusted R-squared** |
|---|---|---|
| Intensity | 0.432 | - |
| OCUI | 0.433 | 0.429 |
| XCUI | 0.435 | 0.439 |
| OHET | 0.498 | 0.491 |
| AHET | 0.470 | 0.473 |
| XHET | 0.503 | 0.512 |
| VHETB | 0.493 | 0.489 |
| VHETF | 0.487 | 0.485 |

(note a result for Intensity using the 2$^{nd}$ Order Approx. form cannot be calculated due to perfect colinearity).

It can be seen that the HET based variables have noticeably higher adjusted R-squared scores than the CUI based ones. The lowest HET adjusted R-squared score is for AHET (i.e. the minimum gaps at the end of each section). This suggests that the size of the minimum gap wherever it occurs is a more important determinant of reactionary delay than the size of arrival gaps. Of the two 'vulnerable' HET variables, the one where the gap before the 'vulnerable' train is measured (i.e. VHETB) has a higher adjusted R-squared score than the one where the gap following the train is used (i.e. VHETF). This suggests that the impact on 'vulnerable' trains is of greater importance than the impact they have on other trains in the resulting level of reactionary delay. However, OHET, which makes no distinction between the type of gaps, has an adjusted R-squared result greater than the two 'vulnerable' measures. The difference in adjusted R-squared is not substantial. It is not believed therefore that OHET is necessarily superior to the two 'vulnerable' measures. Instead it is felt that the latter two measures would benefit from further work on their admittedly rather crude weighting of the gaps between 'vulnerable' trains. Finally, XHET with its inclusion of junction crossing moves has the highest adjusted R-squared result.

In summary therefore, using  two different functional forms the HET based measures act as better predictors of Congested Related Reactionary Delay than the CUI based measures, with XHET (which includes junction crossing moves) performing 'best' of all.

In terms of the most appropriate functional form, as previously noted, the adjusted R-squared values suggest that the 2$^{nd}$ Order Approximation (logarithmic) functional form is the 'best' of the two options to describe the data set. However, the Exponential functional form has been chosen as the 'preferred' option for a number of reasons:-

- Since it has only one explanatory coefficient rather than two it is the more parsimonious of the two functional forms.
- The choice of an Exponential form is in line with the findings of the previous work on the Capacity Charge and the conclusions from other research.
- The Exponential Functional Form still produces adjusted R-squared results that are more than reasonable and fairly close to those produced by the 2$^{nd}$ Order Approximation (Logarithmic) Form.

Furthermore, as noted in Section 5.2.6 (p102) of this thesis Kennedy (2008, p89) warns against the danger of solely relying on the highest adjusted R-squared due to the possibility that specific peculiarities of the data set have contributed to the result rather than that the true underlying relationship has been found. Consideration has therefore been given to why the Exponential form does not perform quite 'as well' as the Second Order Approximation (logarithmic) Form when using the sectional data set.

Comparison of the shape of the Exponential and Second Order Approximation (logarithmic) functional forms, as shown in Figure 7.1. is revealing. This uses the regression output for Welwyn Viaduct (Up) but the other geographic sections produce very similar results.

It can be seen that both functional forms show a similar rate of increase in reactionary delay up until 70 to 80% capacity utilisation. After this point the rate of increase in reactionary delay is significantly greater with the Exponential functional form than for the Second Order Approximation (logarithmic) functional form. The former more closely matches expectations that with high levels of traffic the rate of increase in delay will rise substantially as there are more opportunities for reactionary delay to be incurred. The impact of different sensitivity tests on the calculated adjusted R-squared results will be discussed in the next section.

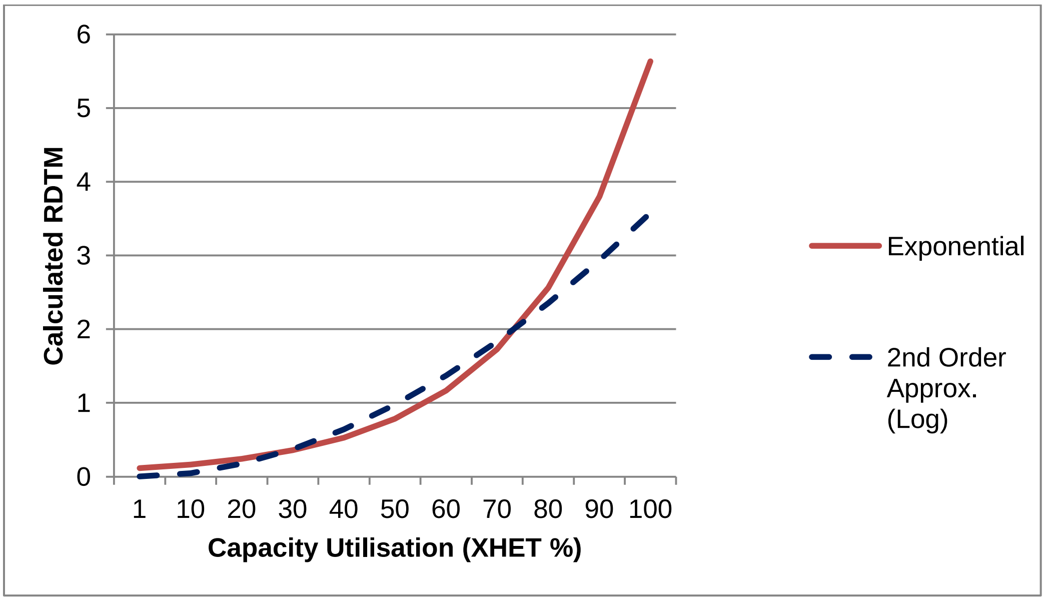


**Figure 7.1** Comparison of the Two Logarithmic Function Forms using the Fixed Effects One Way model and the XHET capacity utilisation measure (Welwyn Viaduct Up Direction).

Figure 7.2 shows the Actual versus Fitted lines for the Exponential and Second Order Approximation (logarithmic) functional forms for the XHET capacity utilisation measure. This is for the Grantham to Newark section although a similar relationship can be observed in the other sections. It can be seen that the two Fitted lines are very similar to each other. Both show differences from the Actual line during the same periods of the day. These equate to some of the highest and lowest levels of observed reactionary delay. However, it can be seen that in contrast to the 2$^{nd}$ Order Approximation (logarithmic) Fitted line, the fitted Exponential line is almost identical to the Actual line for the 1600 to 1700 hour period. This period has the highest calculated capacity utilisation during the day (80.5%). This therefore reinforces the believe that the Exponential functional form is more accurate at predicting the level of reactionary delay at the highest levels of capacity utilisation.

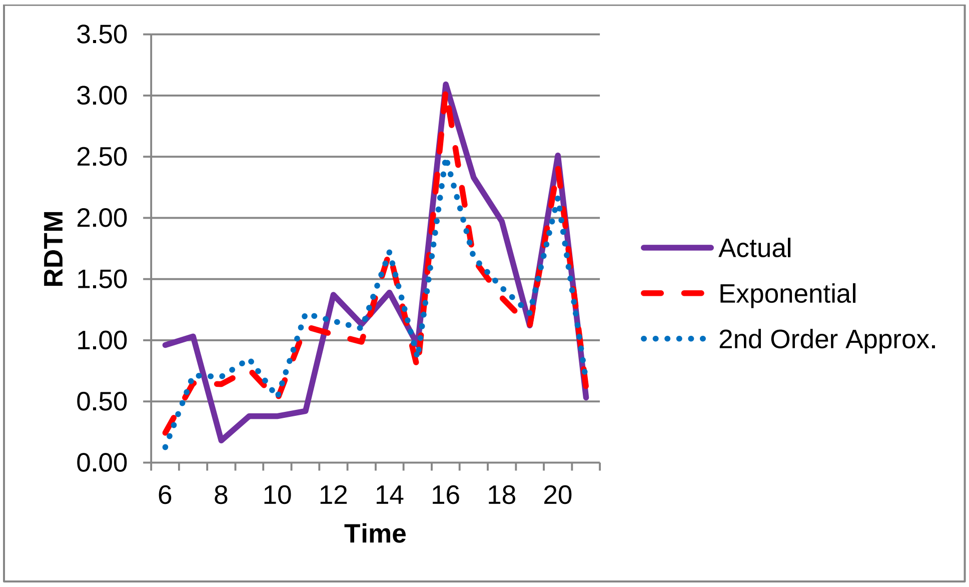


**Figure 7.2** Actual versus Fitted Lines for the Exponential and Second Order Approximation (Logarithmic) Functional Forms for the Grantham to Newark Section and the XHET Capacity Utilisation Measure

### 7.2.6 Calculated Elasticities for Different Sectional Capacity Utilisation Measures and Functional Forms

Table 7.8 shows average elasticities for a sample of the capacity utilisation measures. The preferred functional form (Exponential) and approach (one-way, fixed effects) has been used in each case. In addition results for the Second Order Approximation (logarithmic) functional form are given.

The table shows that for both functional forms elasticities are greater for the HET capacity utilisation measures than the CUI based ones (i.e. a 1% increase in HET capacity utilisation produces a higher percentage increase in reactionary delay). It can also be seen that the Intensity measure, where a relationship can be calculated (i.e. using the Exponential functional form), has a greater elasticity than the CUI measures but not the HET measures. This suggests that a change in the level of traffic spacing in particular has a greater impact on changes in reactionary delay than the actual volume of trains (Intensity) and the volume of capacity used (CUI).

**Table 7.8** Elasticities for a Sample of the Sectional Explanatory Variables from the Two Non-Linear Functional Forms (One-Way / Fixed Effects).

| **Capacity Utilisation Measure** | **Exponential** | **2nd Order Approx. (Logarithmic)** |
|---|---|---|
| Intensity | 0.0299 | - |
| OCUI | 0.0238 | 0.0202 |
| XCUI | 0.0245 | 0.0188 |
| OHET | 0.0383 | 0.0469 |
| XHET | 0.0393 | 0.0335 |

(note a result for Intensity using the 2nd Order Approx. form cannot be calculated due to perfect colinearity).

A particular point of interest with Table 7.8 is that for the Second Order Approximation (Logarithmic) functional form the link-only measures (i.e. OCUI and OHET) produce higher elasticities than the junction and link measures (i.e. XCUI and XHET). This is particularly true for OHET which has a substantially higher elasticity than XHET. It can be seen that the opposite is true for the Exponential functional form. Based on expectations XHET and XCUI should have higher elasticities than their equivalents. This is due to the belief that capacity utilisation at junctions contribute significantly to the overall levels of reactionary delay. This observation further reinforces the view that the Exponential functional form is preferred.

### 7.2.7 Sectional Data Checking

Due to the limited number of data cells used in the regression analysis (384) it was advisable to undertake a degree of data checking. Clearly, outliers and clusters of data points can have a significant impact on the results and the choice of the preferred approach. Figure 7.3 shows each of the points in the data set plotted by calculated XHET capacity utilisation against observed reactionary delay per train mile. Note, the decision has been taken not to plot regression curves for the various functional forms. This is because the use of panel data means that applying a single regression curve to the aggregated data would be fairly meaningless.

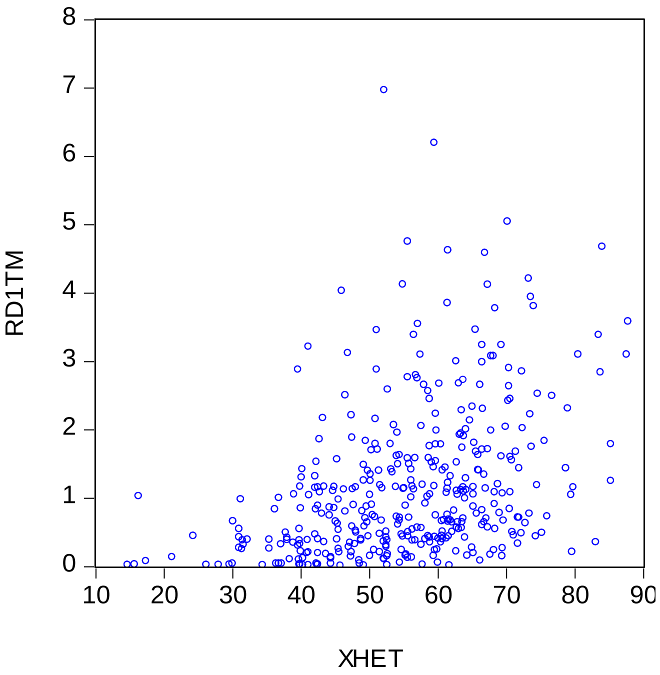


**Figure 7.3** Plotted Data Points for the Sectional Data Set (% XHET compared with Minutes Reactionary Delay per Train Mile).

It can be seen that there are only a small number of outliers and very little data clustering. Low levels of reactionary delay are generally associated with low levels of capacity utilisation and higher levels of delay with higher levels of capacity utilisation. This suggests that the data set is robust.

However, in order to replicate the process carried out for the recalibration of the Capacity Charge, a number of sensitivity tests were carried out. These involved the exclusion of cells where the there was no record of any reactionary delay and the exclusion of cells for various ranges of calculated capacity utilisation (using the OCUI measure). The purpose of this was to examine the sensitivity of the results previously described to changes to the core data set.

Table 7.9 shows the results of the data ‘cleaning’ for a number of capacity utilisation measures and the two functional forms being considered. In each case a fixed effects, one-way approach has been adopted. The removal of cells with no observed reactionary delay clearly produces a marked reduction in the ‘goodness-of-fit’ in all cases. It had been thought that because theoretically zero reactionary delay can be associated with even relatively high levels of capacity utilisation, removal of cells with zero delay could increase the adjusted R-squared results. However, the cells with zero reactionary delay in the sectional data set are all associated with low levels of capacity utilisation. Their removal therefore reduces the calculated relationship between low levels of capacity utilisation and low levels of reactionary delay.

**Table 7.9** Adjusted R-squared Results for Different Measures and Functional Forms Following the 'Cleaning' of the Sectional Data Set.

| **Capacity Utilisation Measure / Functional Form** | **Original Data Set** | **0 RDTM Cells Removed** | **<30% OCUI Cells Removed** | **<30% >75% OCUI Cells Removed** | **>75% OCUI Cells Removed** |
|---|---|---|---|---|---|
| Observations | 384 | 359 | 326 | 319 | 377 |
| OCUI Exponential | 0.433 | 0.399 | 0.421 | 0.398 | 0.439 |
| OCUI 2nd Order Approx (Log) | 0.429 | 0.402 | 0.412 | 0.425 | 0.431 |
| XCUI Exponential | 0.435 | 0.403 | 0.398 | 0.398 | 0.442 |
| XCUI 2nd Order Approx (Log) | 0.439 | 0.405 | 0.448 | 0.427 | 0.442 |
| OHET Exponential | 0.498 | 0.435 | 0.424 | 0.421 | 0.498 |
| OHET 2nd Order Approx (Log) | 0.491 | 0.435 | 0.468 | 0.470 | 0.490 |
| XHET Exponential | 0.503 | 0.453 | 0.479 | 0.417 | 0.504 |
| XHET 2nd Order Approx (Log) | 0.512 | 0.451 | 0.465 | 0.466 | 0.512 |

In terms of the capacity utilisation ranges, only when cells with a calculated OCUI of greater than 75% are removed are slightly better adjusted R-squared results achieved. This suggests that the capacity utilisation measures are slightly less accurate at very high levels of capacity utilisation. This is possibly due to the small number of cells (seven) in the data set with such a high degree of congestion. Both the other ranges of capacity utilisation generally show a marked reduction in the level of the adjusted R-squared result. A noticeable exception is XCUI with a Second Order Approximation (logarithmic) functional form where the removal of cells with calculated OCUI capacity below 30% produces the highest adjusted R-squared result for all the CUI based results. Despite this exception it is clear though that all the different levels of capacity utilisation observed in the data set contribute to the relatively high levels of adjusted R-squared observed.

One further interesting point to make though is how the relationship between the Exponential and Second Order Approximation (logarithmic) results change with different sensitivity tests. For example, with the original results the Second Order Approximation form produces a higher adjusted R-squared result than the Exponential functional form. However, following the removal of any cells with zero reactionary delay the situation is reversed. The results shown in Table 7.9 therefore underlines the decision that the preferred functional form should not be chosen solely based on adjusted R-squared results.

In summary therefore, the Exponential functional form using a one-way, Fixed effects approach is preferred for describing the relationship between capacity utilisation and reactionary delay. As described in Chapter Four this matches the conclusions of the recalibration of the Capacity Charge which took place in 2013 (Arup, 2013). However, as outlined here the alternative ‘HET’ based measures have been found to provide a better indication of reactionary delay than the ‘CUI’ based measures. In particular XHET, which takes into account the size of minimum gaps on both links and at junctions, has been found to the most effective capacity measure of all the ones considered.

The reasons behind the effectiveness of HET compared with CUI at predicting the levels of reactionary delay in the data set will be discussed later in the Chapter. The detailed results themselves are presented in Appendix C.

### 7.2.8 The Inclusion of ‘Other’ Explanatory Variables in the Regression Analysis

The next step was to consider whether the addition to the equations of the ‘other’ explanatory variables described in Chapter Three[38] could improve their accuracy. This was in response to the possibility of omitted variable bias.

As outlined in Chapter Three these variables are divided into two types. Firstly, those that attempt to account for bias due to network effects (i.e. the Section Before, Section After and Time Period Before variables and the Average Distance Travelled and Average Entry Lateness variables). Secondly, there are those variables which are intended to add to the explanatory power of the capacity variables within each of the geographic sections (i.e. the Timetable Complexity, Average Transit Time Variance and Stability variables).

The ‘other’ variables were included in the specification through first including all of them with each capacity variable. Those that were found to be significant were then ‘rerun’ with the relevant capacity variable. In addition, the non-capacity utilisation measures were regressed individually.

Tables 7.10 and 7.11 show the results of the regression analyses which include these additional variables. Since the intention was to determine whether the inclusion of extra variables could substantially increase the explanatory power of each equation, the original results are also shown for each of the capacity variables. The analyses were carried out for the Exponential functional form using a one-way, fixed effects approach (established in the previous sections as being the preferred model).

Tests were again carried out for evidence of autocorrelation and heteroskedasticity in the sectional capacity variables. Once again although there was no evidence of autocorrelation or measurement error, heteroskedasticity was detected. This was again accounted for using the White Heteroskedasticity Consistent Covariance Matrix Estimator approach.

[38] See Table 3.5

**Table 7.10** Comparison of t-statistic Scores (Heteroskedasticity Adjusted) Following the addition of 'Other' Explanatory Variables to the Specifications.

| Capacity Variable | Original Capacity t-statistic (White) | New Capacity t-statistic (White) | 'Other' Variable t-statistic (White) | 'Other' Variable t-statistic (White) | 'Other' Variable t-statistic (White) |
|---|---|---|---|---|---|
| Intensity | 4.314 | -1.002 | TTC<br>3.789 | TBCAP<br>3.290 | SFCAP<br>3.113 |
| OCUI | 4.378 | -0.865 | TTC<br>3.702 | TBCAP<br>2.282 | SFCAP<br>3.751 |
| XCUI | 4.597 | 0.010 | TTC<br>3.844 | TBCAP<br>2.647 | SFCAP<br>3.954 |
| OHET | 7.931 | 7.931 | - | - | - |
| AHET | 6.344 | 4.902 | TTC<br>3.174 | - | - |
| XHET | 8.601 | 4.897 | - | - | SFCAP<br>3.081 |
| VHETB | 7.574 | 7.574 | - | - | - |
| VHETF | 7.392 | 6.048 | TTC<br>2.176 |  | - |

(For key to the 'Other' Variables shown in the Table please see Abbreviations section at the end of this thesis).

Table 7.10 shows those additional variables that were found to be significant when included with each of the capacity variables. It can be seen that of the eight 'other' variables being considered only half of them were found to be significant when included in one or more of the regression specifications.

Of the non-capacity utilisation related measures, only Timetable Complexity (TTC) was found to be significant when combined with the explanatory variables already being considered. It can be seen that it ‘added value’ to six of the eight regressions.

Average Distance Travelled (ADT) and  Average Entry Lateness (AEL) can be said to represent the ‘situation’ in the network prior to the section in question. Their lack of significance suggests this situation does not have a substantial bearing on the level of reactionary delay experienced. As described previously these two have been used in other research. As noted in Chapter Three, Entry Lateness was not found to be a significant determinant of reactionary delay by A.Lindfeldt (2012). In contrast Average Distance Travelled has been found to be significant (as noted by Olsson and Haugland, 2004), however this was as a determinant of punctuality at final destination rather than delays en-route and for long-distance passenger trains rather than all traffic.

Capacity utilisation in the section before (i.e.SBCAP) is also not significant with any of the original capacity utilisation measures. Its lack of significance combined with those of ADT and AEL suggests that the situation in advance of a section is not an important factor in determining its level of reactionary delay.

The lack of significance of the Stability variable when combined with capacity utilisation measures is perhaps surprising. However, it could be argued that the role of allowances is to reduce the overall level of delay and in particular its further propagation, rather than simply delays in the geographic section the allowances are located. Their use of capacity (as shown by the Capacity Balance diagram reproduced in Chapter Three) also means that they are used sparingly.

The lack of significance of Average Transit Time Variation (ATV) is likely to reflect the fact that the impact of variations in journey time is accounted for in the majority of capacity utilisation measures i.e. in the form of heterogeneity.

Table 7.10 shows that along with Timetable Complexity, capacity utilisation in the time period before and on the section following (i.e. TBCAP and SFCAP) are significant ‘other variables’. However, it is important to note that Intensity and the two CUI variables actually become insignificant themselves following the addition of these ‘other’ variables.

SFCAP is also significant when combined with XHET. Its significance when combined with four of the original capacity utilisation variables suggests that congestion following a location is of greater importance in determining the level of reactionary delay. This observation is reinforced by the lack of significance of ADT, AEL and SBCAP.

Another interesting comment that cannot be made about the results shown in Table 7.10 is that the relationships between both OHET and VHETB and reactionary delay do not benefit from the addition of any of the eight ‘other’ variables to the specification.

Finally, it was found that when the five non-capacity utilisation variables were analysed as a separate group only Timetable Complexity was found to be significant. However, when each of them was looked at completely on their own each of them was found to be significant. This suggests that factors other than capacity utilisation have an important impact on the level of reactionary delay. However, the complexity of the timetable is clearly a key determinant of reactionary delay.

The addition of other variables can therefore increase the accuracy of the relationship between capacity utilisation and reactionary delay. However, it is necessary to note the size of this improvement. This can be achieved by comparing the original adjusted R-squared values with the new ones for each capacity variable. Table 7.11 compares the adjusted R-squared results for the capacity utilisation variables with those achieved following the addition of the ‘other’ variables.

Table 7.11 shows that the adjusted R-squared for Intensity, OCUI and XCUI substantially increase following the addition of ‘other’ variables. However, as noted previously in each case the capacity utilisation measure itself ceases to be significant. Therefore, for the Intensity and CUI approaches the measured congestion in adjacent sections and in the time period before, as well as the complexity of the timetable in the section itself, describe the relationship between utilisation and reactionary delay better than using the measured level of capacity utilisation for the section itself.

In contrast, the HET based measures show a smaller increase in the adjusted R-squared scores where ‘other’ variables are significant. The biggest increase occurs with the addition of ‘Section Following’ measure to XHET (from 0.503 to 0.517). This is also the highest adjusted R-squared result produced by the analysis.

**Table 7.11** Comparison of Adjusted R-squared Scores Following the Inclusion of 'Other' Explanatory Variables

| Capacity Utilisation Variable | Original Adjusted R-squared | New Adjusted R-squared |
|---|---|---|
| Intensity | 0.432 | 0.484 |
| OCUI | 0.433 | 0.492 |
| XCUI | 0.435 | 0.498 |
| OHET | 0.498 | 0.498 |
| AHET | 0.470 | 0.482 |
| XHET | 0.503 | 0.517 |
| VHETB | 0.493 | 0.493 |
| VHETF | 0.487 | 0.493 |
| TTC | n/a | 0.442 |
| ADT | n/a | 0.405 |
| AEL | n/a | 0.417 |
| ATV | n/a | 0.398 |
| STAB | n/a | 0.401 |

(For key to 'Other' Variables please see Abbreviations section at the end of this thesis).

Finally, Table 7.11 shows that using 'other' variables alone also gives fairly reasonable adjusted R-squared. In the case of Timetable Complexity it is worth noting that this explanatory variable has a higher adjusted R-squared than the Intensity or CUI based variables. In contrast the four 'other' variables produce the lowest adjusted R-squared values suggesting that their value as determinants of reactionary delay is not as great as any of the capacity utilisation measures studied.

Table 7.12 shows the correlation between the sectional capacity variables and the significant 'other' variables. It can be seen that there is a strong correlation between a number of capacity variables (Intensity, OCUI, AHET and VHETF ) and their significant 'other' variables. There is therefore a substantial amount of overlap between capacity utilisation and 'other' explanatory variables. This underlines the fact that there are a number of complex relationships between the factors which cause reactionary delay. Table 7.12 also shows that weaker correlations exist. For example, the correlation between XCUI and Timetable Complexity is 0.37. However, in the case of XCUI, three 'other' variables are significant and contribute to the raised adjusted R-squared value shown in Table 7.11.

**Table 7.12** Matrix Showing the Correlation Between Sectional Capacity Variables and 'Other' Explanatory Variables

| **Capacity Utilisation Variable** | **'Other' Variable** | **'Other' Variable** | **'Other' Variable** |
|---|---|---|---|
| Intensity | TTC<br>0.74 | TBCAP<br>0.71 | SFCAP<br>0.51 |
| OCUI | TTC<br>0.64 | TBCAP<br>0.59 | SFCAP<br>0.58 |
| XCUI | TTC<br>0.37 | TBCAP<br>0.54 | SFCAP<br>0.55 |
| OHET | - | - | - |
| AHET | TTC<br>0.71 | - | - |
| XHET | - | - | SFCAP<br>0.45 |
| VHETB | - | - | - |
| VHETF | TTC<br>0.68 | - | - |

It can be concluded that the addition of other variables does 'improve' the relationship between capacity utilisation and reactionary delay. However, the level of 'improvement' seen is variable as are the actual 'other' variables that are significant. In the case of Intensity, OCUI and XCUI a number of 'other' variables acting in combination actually produce better results than using these capacity utilisation measures alone. Furthermore, use of the Timetable Complexity variable on its own also produces a 'better' result. This further reinforces the view that although capacity utilisation measured using the CUI approach is a significant determinant of reactionary delay there are more effective and appropriate approaches.

Despite the results it has been decided not to pursue the addition of ‘other’ variables. This is because they add to the complexity of the regression equations .In the case of the Intensity and CUI based equations although there is a substantial increase in the adjusted R-squared score this is at the expense of the addition of multiple ‘other’ variables and the capacity utilisation variables themselves becoming insignificant. For the HET based measures it is felt the adjusted R-squared value do not increase substantially enough to warrant the addition of another variable to the equation. As noted in Chapter One, one objective of this thesis was to consider the transferability of the results. A specification with only one explanatory variable is considered much more transferrable than one with several explanatory variables, whilst remaining consistent with the approach adopted for the Capacity Charge.

A final point to make concerns the value of β. As discussed XHET is the preferred capacity utilisation measure. The only ‘other’ variable that is significant with it is ‘Section Following Capacity’ (i.e. SFCap). The value of β for XHET on its own is 0.039394. XHET and SFCap together are 0.027891 and 0.018523 respectively. This suggests that the use of a fixed effects approach is helping to account for the possibility of omitted variable bias and network effects by producing a β similar to the combined values of XHET and SFCap.

### 7.2.9 Different Dependent and Explanatory Variables

Two final aspects need to be considered to complete the analysis of the sectional data set. Firstly, the suitability of the dependent variable used in the analysis of the data set needs to be considered. Secondly, the results of this analysis can then be used to inform the examination of alternative explanatory variables to the capacity utilisation and ‘other’ variables already discussed in this thesis.

As described previously the choice of the dependent variable used in this thesis was the one chosen for the calculation of the Capacity Charge. Faber Maunsell note that the choice of CRRD per train mile[39] represents a more

[39] Or more accurately (CRRD+1) / Train Miles to allow for the use of logs in the functional forms.

than linearly increase in reactionary delay and thus shows any increased marginal cost due to congestion (Faber Maunsell 2007, p9).

It is however appropriate to consider the impact on the results of a number of alternative dependent variables. Two alternative dependent variables have been considered. These are firstly, simply the CRRD+1 per cell (RD1) and secondly, (CRRD+1 ) / mileage (RD1M). The results are shown in Table 7.13 for XCUI and XHET[40]. To maintain consistency with the preferred approach the results are for the Exponential functional form and one-way, fixed effects.

**Table 7.13** Comparison of Adjusted R-squared Results for the Original and Two Alternative Dependent Variables (Exponential , One-Way, Fixed Effects).

| **Capacity Utilisation Measure / Statistic** | **Original (RD1TM)** | **New 1 (RD1)** | **New 2 (RD1M)** |
|---|---|---|---|
| XCUI<br>Adjusted R-squared | 0.438 | 0.400 | 0.552 |
| XCUI<br>t-statistic<br>(Heteroskedasticity) | 3.596 | 7.745 | 7.759 |
| XHET<br>Adjusted R-squared | 0.503 | 0.488 | 0.614 |
| XHET<br>t-statistic<br>(Heteroskedasticity) | 8.482 | 12.138 | 11.977 |

The t-statistic results shown in the table reveal that both of the new explanatory factors remain significant following the adoption of the two alternative dependent variables. This is as expected due to the clear link that has already been established between capacity utilisation and reactionary delay. Note, that in line with the previous findings discussed in the thesis these have been adjusted to account for heteroskedasticity.

[40] i.e. the two ‘best’ performing CUI and HET capacity utilisation variables.

Overall, the adjusted R-squared results are unsurprising. They show a better relationship between reactionary delay per mile and capacity utilisation than when the reactionary delay has been divided by train numbers. This is because the expected relationship that high levels of capacity utilisation is equated with high levels of reactionary delay and low levels of capacity utilisation means low levels of delay is strengthened. However, the length of the section over which traffic experiences congestion is also clearly important. This is suggested by the poorer performance of RD1 where reactionary delay has not been divided by either section length or traffic numbers. Possibly this is because a long geographic section with a low level of capacity utilisation could arguably produce a similar level of reactionary delay to a short section with high utilisation simply due to the increased time that traffic is exposed to any performance issues.

The results show that RD1TM (i.e. (CRRD+1) / Train Miles) is a more appropriate dependent variable to use than RD1 (i.e. CRRD+1). Although, the dependent variable RD1M (i.e. not divided by train numbers) performs better than RD1TM, the latter is still preferred. This is because as noted by Faber Maunsell (2007, p9) the intention of the dependent variable was to reflect the any increases in marginal cost due to rising congestion. The adoption of this approach is as discussed in Chapter 4 of this thesis in line with the principles of congestion charging. Dividing reactionary delay per mile by train numbers removes the average reactionary delay (i.e. Average Cost) per train from the charging regime.

Table 7.14 shows the two new alternative explanatory variables considered at this point. Once again for the sake of consistency the fixed effects, one-way approach has been adopted and the two non-linear functional forms have been used. Only the dependent variable RD1 (i.e. CRRD+1 per cell) has been used due to the high correlation of these explanatory variables with  the length and traffic numbers of each cell. Once again the t-statistic results have been adjusted to account for Heteroskedasticity.

The t-statistic for the Mileage variable in Table 7.14 is not significant. This shows that by itself length of exposure to possible performance risk is insufficient to explain the level of reactionary delay observed. It does perhaps suggest that an explanatory variable used by itself does have to reflect in some way the utilisation of the rail network. This view is supported by the results for the Train Miles variable which is both significant and has an adjusted R-squared value similar to those calculated for some of the capacity utilisation variables used in this thesis (see Table 7.7).

**Table 7.14** Comparison of Adjusted R-squared Results for the Two Alternative Explanatory Variables (Exponential, One-Way, Fixed Effects)

| Explanatory Variable | RD1 Dependent Variable |
|---|---|
| Mileage<br>Adjusted R-sq | 0.267 |
| Mileage<br>t-statistic<br>(Heteroskedasticity) | -1.823 |
| Train Miles<br>Adjusted R-sq | 0.429 |
| Train Miles<br>t-statistic<br>(Heteroskedasticity) | 7.950 |

One final useful piece of analysis is to examine the impact of a specification with a dependent variable of Reactionary Delay (RD1) and the combination of XHET and Train Miles as the explanatory variables. Using the Exponential, one-way, fixed effects approach, both explanatory variables are significant and the calculated adjusted R-squared result is 0.500. Given that the adjusted R-square for XHET for the standard dependent variable is 0.503 (see Table 7.7); there is clearly nothing to be gained from adopting this approach.

### 7.2.10 Conclusions

In conclusion, the XHET variable using an Exponential form and a one-way, ‘fixed’ effects model is considered to be the preferred approach to predicting the level of reactionary delay on the sample network. Additionally, it is believed the dependent variable used in the analysis (i.e. CCRD+1/Train Miles) is the most appropriate for understanding the implications of the findings for the pricing of congested rail networks.

The implications of this for understanding the actual relationship between capacity utilisation and reactionary delay are considered in the next section. This presents and discusses the β values obtained for each of the sectional capacity utilisation variables. Since β is the slope parameter, the size of this indicates for each variable how much greater than linearly reactionary delay increases as the sample network becomes more congested.

## 7.3 β Values for the Sectional Capacity Utilisation Measures

Table 7.15 shows the calculated route specific coefficient (i.e. β) for each of the sectional capacity variables for the Exponential functional form for a one-way ‘fixed’ effects approach. It will be seen that these are obviously the elasticities already presented in Table 7.8. However, in this case the βs are given for each of the Sectional Capacity Utilisation variables. This allows a more detailed comparison to take place.

**Table 7.15** Calculated β’s for the Sectional Capacity Variables (Exponential Form with a One-Way ‘Fixed’ Effects Approach).

| Variable | Route | Calculated β |
|---|---|---|
| Intensity | Part ECML | 0.0299 |
| OCUI | Part ECML | 0.0238 |
| XCUI | Part ECML | 0.0245 |
| OHET | Part ECML | 0.0383 |
| AHET | Part ECML | 0.0368 |
| XHET | Part ECML | 0.0393 |
| VHETB | Part ECML | 0.0353 |
| VHETF | Part ECML | 0.0349 |

A number of observations can be made about the contents of Table 7.15[41]:-

- the calculated β’s, are substantially higher for the HET based capacity variables than either the Intensity or CUI based variables.
- the two ‘junction’ variables both produce steeper curves than the alternate ‘link-only’ variables.

[41] It has already been discussed in Chapter Three how CUI and HET can be directly compared. This is due to the common use of planning headways (and margins) which is also shared by the Intensity Variable.

- the Intensity Variable produces a steeper curve than either of the two CUI variables.
- the preferred variable (XHET) produces the steepest curve of all the capacity variables considered.
- The closet variable to the approach adopted for the calculation of the Capacity Charge (OCUI) produces the shallowest curve of all the capacity variables considered.

In simple terms therefore the more effective an estimator of reactionary delay the 'steeper' the curve.

Of course the specifications also include individual '*A*' coefficient values for each geographic section. These are presented for each capacity utilisation variable in Appendix C. It is worth noting here though, that the values for each of these vary considerably between the individual sections.  The variation in the size of these 'dummy' variables indicates that each individual section has its own unique impact on the associated amounts of reactionary delay. It is also important to note that the more 'effective' HET measures have lower '*A*' values than the CUI based and Intensity capacity measures.

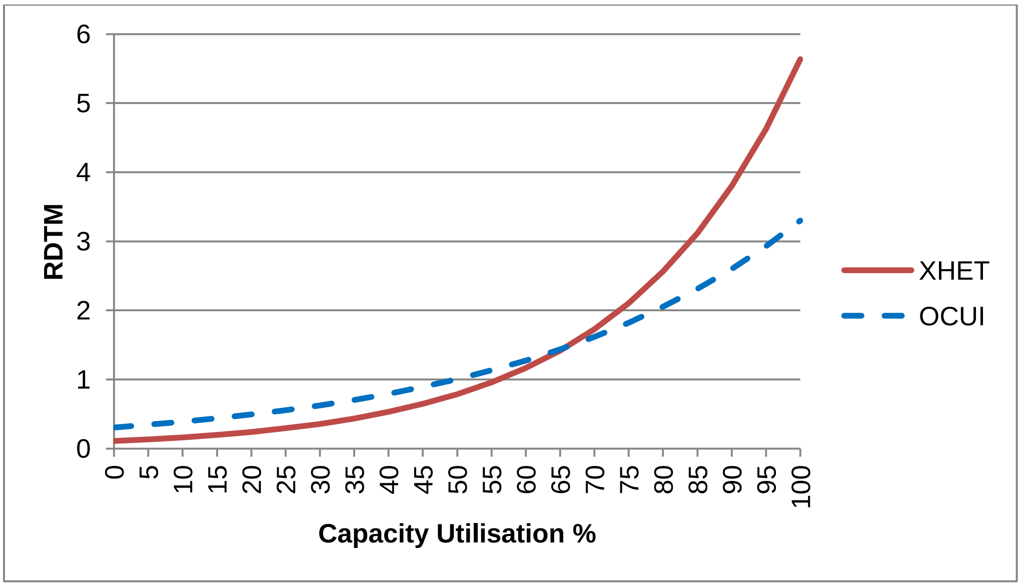


**Figure 7.4** OCUI and XHET Regression lines for the Grantham to Newark Section.

This leads to noticeably different regression lines for the alternate types of capacity utilisation measures. This is illustrated in Figure 7.4 which shows the regression lines for OCUI and XHET for the Grantham to Newark Section.

XHET's steeper curve can clearly be seen. At high capacity utilisation level sit predicts considerably more reactionary delay than OCUI. However, it is also noticeable that due to its flatter curve the OCUI relationship predicts higher levels of reactionary delay at low levels of capacity utilisation.

## 7.4 Regression Results for the Area Capacity Measures

### 7.4.1 Identification of the most appropriate functional form

Once again a Box-Cox transformation of the data was employed using the method described by Dougherty (2011). For reasons of brevity Table 7.16 only shows the calculated residual sums of squares for the fixed effects one-way approach. The other results are shown in Appendix B and mirror the conclusions reached here.

It can be seen that no results are given for the Second Order Approximation (logarithmic) functional form. This is because, as noted, due to the presence of perfect colinearity it is not possible to calculate the results of a regression using this functional form.

**Table 7.16** Residual Sums of Squares for the Five Functional Forms (One-Way / Fixed Effects) for the Area Variables.

| **Capacity Utilisation Variable** | **Linear (linear)** | **Quadratic (linear)** | **2nd Order Approx. (linear)** | **Exponential (logarithmic)** | **2nd Order Approx. (logarithmic)** |
|---|---|---|---|---|---|
| LHET | 12.195 | 12.326 | 12.194 | 15.155 | - |
| LCUI | 11.988 | 12.455 | 11.237 | 15.212 | - |
| EHET | 10.105 | 10.289 | 9.979 | 14.128 | - |

(note a result for Intensity using the 2nd Order Approx. form cannot be calculated due to perfect colinearity).

Table 7.16 shows that in every case  the residual sums of the squares are lower for the linear based functional forms than the Exponential form. Although, contrary to expectations and the findings for the sectional explanatory variables, the decision was therefore taken to proceed with the linear  functional forms.

### 7.4.2 Fixed and Random Effects for the Area Variables

Table 7.17 shows the results of the Hausman Tests for the Area capacity utilisation measures.

**Table 7.17** Chi Square Statistics Calculated by the Hausman Test for the Area Variables

| Capacity Utilisation Variable | One-Way Linear | Two-Way Linear | One-Way Quadratic | Two-Way Quadratic | One-Way 2nd Order Approx. (Linear) | Two-Way 2nd Order Approx (Linear) |
|---|---|---|---|---|---|---|
| LCUI | 0.001 | 0.000 | 0.051 | 0.103 | 0.285 | 0.381 |
| LHET | 0.489 | 0.004 | 0.377 | 0.006 | 3.727 | 0.169 |
| EHET | 0.311 | 0.898 | 0.250 | 1.054 | 9.428 | 1.612 |

The critical value for the Exponential functional form is 3.815 at a 95% Confidence Interval. The critical value for the Second Order Approximation (logarithmic) form is 5.992. It can therefore be seen that with the exception of the one-way approach for the Second Order Approximation (linear) functional form for EHET, the null hypothesis cannot be rejected and a random effects approach is recommended. However, for the reasons stated on page 154 of this thesis the decision has been taken to proceed with a fixed effects approach.

### 7.4.3 One-Way and Two-Way Models for the Area Explanatory Variables

Table 7.18 shows the t-statistic results for the one and two-way approaches for the Linear and Quadratic functional forms. Table 7.19 shows the F-tests of Joint Significance results for the Second Order Approximation (linear) functional form. Table 7.20 shows the adjusted R-squared results for each of the three functional forms. In line with the findings described in Section 7.4.2, all results are for Fixed effects.

The t-statistic critical value for the size of data set with a 95% Confidence Interval is 1.999. It can be seen that for all capacity utilisation variables, the t-statistic scores are significant for the one-way approach. However, LHET (Linear and Quadratic) and LCUI (Quadratic) are not significant assuming the two-way approach.

**Table 7.18** T-statistic Results for the One-Way and Two-Way Models for the Area Variables (Linear and Quadratic Functional Forms).

| Capacity Utilisation Variable | Linear One-Way FE | Linear Two–Way FE | Quadratic One-Way FE | Quadratic Two–Way FE |
|---|---|---|---|---|
| LCUI | 2.987 | 2.230 | 2.525 | 1.798 |
| LHET | 2.787 | 1.865 | 2.657 | 1.883 |
| EHET | 4.645 | 3.276 | 4.487 | 3.121 |

The F-test results are shown in Table 7.19.The critical value for a 95% Confidence Interval for the size of data set is 3.148. This means that the LHET capacity utilisation measure for the two-way approach is not significant using this functional form.

**Table 7.19** F-Test of Joint Significance Results for the One-Way and Two-Way approaches (Second Order Approximation (Linear)).

| Capacity Utilisation Variable | 2nd Order Approx. (Linear) One Way | 2nd Order Approx. (Linear) Two-Way |
|---|---|---|
| LCUI | 7.071 | 6.365 |
| LHET | 4.083 | 2.529 |
| EHET | 11.867 | 8.526 |

Table 7.20 shows the adjusted R-squared results for the one-way and two-way approaches for each of the three functional forms being considered. It can be seen that there is a great deal of variation in the size of the results across the different capacity utilisation variables, functional forms and model approaches (i.e. one-way or two-way). In all cases though it can be seen that the EHET measure has a higher adjusted R-square than the equivalent results for LHET and LCUI.

**Table 7.20** Comparison of Adjusted R-Squared Results for the One-Way and Two-Way Models for the Area Variables.

| **Capacity Utilisation Variable** | **LCUI** | **LHET** | **EHET** |
|---|---|---|---|
| Linear One-Way | 0.174 | 0.160 | 0.304 |
| Linear Two-Way | 0.175 | 0.149 | 0.262 |
| Quadratic One-Way | 0.142 | 0.151 | 0.291 |
| Quadratic Two-Way | 0.145 | 0.151 | 0.249 |
| 2nd Order Approx. (Linear) One-Way | 0.212 | 0.145 | 0.301 |
| 2nd Order Approx. (Linear) Two-Way | 0.221 | 0.132 | 0.263 |

Following the analysis described above the decision was taken to adopt the one-way approach. This was for the following reasons:-

- As discussed previously, a one-way approach is much more intuitive.
- As shown in Tables 7.18 and 7.19, a number of the variables assuming the two-way approach have been found to be insignificant.
- A one-way approach is consistent with the original calibration and recalibration of the Capacity Charge.
- The adoption of a one-way approach is consistent with the decision taken for the sectional explanatory variables as described earlier in this chapter.

It is also worth noting that it has been checked that the favouring of a one-way over a two-way approach does not affect the decision over which capacity variable and functional form is preferred.

This means that the preference is for a one-way fixed effects approach. This therefore replicates the findings for the sectional explanatory variables and the conclusions of the recalibration work in 2013 for the Capacity Charge (Arup, 2013).

### 7.4.4 Tests for Autocorrelation and Heteroskedasticity

Once again autocorrelation and heteroskedasticity were then tested for. In line with the findings for the sectional capacity variables, no evidence of autocorrelation in the area capacity variables was identified. Additionally and in contrast to the sectional capacity variables, there was also no evidence of heteroskedasticity identified. As described for the sectional variables, the presence of heteroskedasticity was tested using a dummy variable approach for the one-way model. This mixture of heteroskedasticity at a sectional level and homoskedasticity is difficult to explain. It may simply reflect a different relationship between reactionary delay to the meso and macro capacity utilisation variables used.

### 7.4.5 Choice Between Functional Forms and Area Explanatory Variables

Table 7.21 shows the adjusted R-squared results for the Fixed effects one-way approach model .

It can be seen that the ‘fit’ of the capacity utilisation / reactionary delay curves (as shown by the adjusted R-squared scores) vary between the three variables and the three functional forms. For every functional form though, EHET produces a substantially greater adjusted R-square result than either LHET or LCUI. However, LCUI produces a higher adjusted R-square than LHET for two out of the three functional forms (i.e. Linear and Second Order Approx. (Linear)).

These results are interesting as the better performance of EHET over the local ‘Theory of Constraints’ measures (LHET and LCUI) indicate that the minimum spacing of trains wherever that occurs is a more effective indicator of overall reactionary delay in a network than the capacity utilisation at its primary constraint. This does have some logic. The ‘flow’ through a constraint, although heavy, might be fairly evenly spaced; compared with a lighter but more ‘bunched’ flow elsewhere. Indeed, the existence of ‘Trigger-Neck’ congestion as proposed by Vickrey (1961) and described in Chapter

Three supports the view that the impact of a constraint on 'flow' is not necessarily limited to its immediate location.

**Table 7.21** Adjusted R-squared Results for the Three Functional Forms for the Three Area Explanatory Variables (Assuming a Fixed Effects, One-Way Approach).

| Capacity Utilisation Variable | Linear Adjusted R-squared | Quadratic Adjusted R-squared | 2nd Order Approx. (Linear) Adjusted R-squared |
|---|---|---|---|
| LCUI | 0.174 | 0.142 | 0.212 |
| LHET | 0.160 | 0.151 | 0.145 |
| EHET | 0.304 | 0.291 | 0.301 |

There is also the case that other constraints in the network will have an impact on reactionary delay. For example, in the case of the Welwyn areas, although Welwyn Viaduct has been used as the 'primary' constraint the existence of Hitchin Cambridge Junction (which in this data set is an 'at-grade' junction) also clearly has an important impact on reactionary delay. This point is clearly demonstrated by the better performance of the Junction and Link based sectional capacity utilisation measures (i.e. XHET and XCUI) than their equivalent link-only based measures (i.e. OHET and OCUI).

Nonetheless both LHET and LCUI are significant explanatory variables for reactionary delay. The level of capacity utilisation at the primary constraint is therefore an important factor in the overall level of reactionary delay in the surrounding network. The generally better performance of LCUI suggests that the amount of capacity used at primary constraints is more significant than how it is used, in determining the level of reactionary delay in the surrounding network. Once again there is some logic to this. The CUI measurement of the volume of capacity used rather than actually how it is used (as measured by HET) is perhaps more consistent with the Theory of Constraints philosophy, which suggests that the overall flow through a constraint will dictate the performance of the entire network.

The variation in the results seen in Table 7.21 makes the identification of a preferred functional form difficult. The next section provides the calculated elasticities for each area capacity utilisation measure and functional form. This information will be used to assist the decision about the most appropriate functional form to adopt.

### 7.4.6 Calculated Elasticities for the Area Capacity Utilisation Measures

Table 7.22 shows the calculated average elasticities for the area capacity utilisation measures for the three linear functional forms.

The table shows some unexpected results. Firstly, it can be seen that the Second Order Approximation (Linear) functional form produces very low elasticities for the LCUI and EHET area explanatory variables. Examination of the calculated rate of increase in reactionary delay shows that at times this is negative, hence the low average. This is contrary to expectations based on the previous research described earlier in the thesis. For this reason, despite having a good adjusted R-square compared to the other two functional forms; it has been decided not to proceed with this functional form.

**Table 7.22** Calculated Elasticities for the Area Variables (Linear, One-Way / Fixed Effects).

| Capacity Utilisation Measure | Linear Elasticity | Quadratic Elasticity | 2$^{nd}$ Order Approx. (Linear) Elasticity |
|---|---|---|---|
| LCUI | 0.0129 | 0.0117 | 0.0031 |
| LHET | 0.0138 | 0.0112 | 0.0166 |
| EHET | 0.0203 | 0.0103 | 0.0005 |

Secondly, the table shows that EHET has a much higher elasticity than the two Theory of Constraints variables in the case of the Linear functional form; but a slightly lower elasticity in the case of the Quadratic functional form. The former appears more intuitive as it suggests that reactionary delay in an area is more sensitive to changes in traffic bunching within the overall area than due to specific capacity utilisation at its key constraints. It can also be seen that for the Linear functional form LHET has a higher elasticity than LCUI. This matches the findings from the sectional analysis (see Table 7.8). However, the opposite is the case for the Quadratic functional from.

Therefore, of the functional forms considered for the area analysis the Linear produces the ‘better’ results overall. However, the functional form itself does not meet expectations since it implies that the rate of growth in reactionary delay, beyond that implied by an initial linear increase due to train numbers, is static. It is possible that this is simply due to the nature of the data set. The next section therefore outlines the sensitivity tests that were undertaken for the area analysis.

### 7.4.7 Area Data Checking

As with the sectional data set it is necessary to carry out a degree of checking due to the limited number of data points. The substantially fewer number of points (64) and the issues described in the previous section makes this even more important. Figure 7.5 plots EHET against RD1TM.

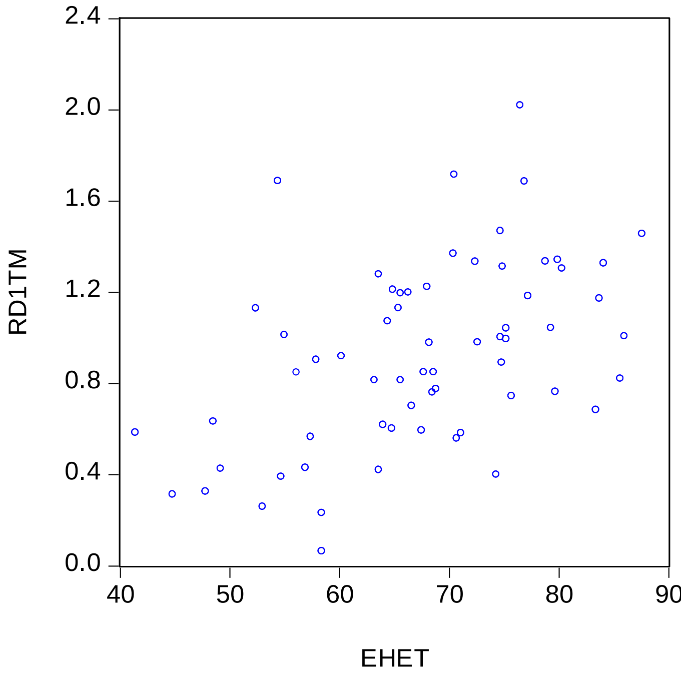


**Figure 7.5** Plotted Data Points for the Area Data Set (% EHET compared with Minutes Reactionary Delay per Train Mile).

It can be seen that there are only a small number of outliers and a limited amount of data clustering. Generally, low levels of reactionary delay are associated with low levels of capacity utilisation and higher levels of delay are associated with increased levels of capacity utilisation. This suggests that the area data set is robust.

However, again the decision was taken to undertake a number of sensitivity tests in line with the approach taken for the recalibration of the Capacity

Charge. This followed the approach adopted for the sectional data set; whereby cells with zero reactionary delay were excluded as were those with a range of capacity utilisation figures. The results are presented in Table 7.23 and are for a fixed effects, one-way approach.

It can be seen that unlike the equivalent sectional analysis there is no column showing the removal of cells with zero reactionary delay from the area data set. This is because all 64 cells had some reactionary delay recorded. Once again though cells with a capacity utilisation of a certain level (this time measured using LCUI) have been removed from the data set.

**Table 7.23** Adjusted R-squared Results for Different Measures and Functional Forms Following the 'Cleaning' of the Sectional Data Set.

| **Capacity Utilisation Measure / Functional Form** | **Original Data Set** | **<30% LCUI Cells Removed** | **<30% >75% LCUI Cells Removed** | **>75% LCUI Cells Removed** |
|---|---|---|---|---|
| LCUI Linear | 0.174 | 0.142 | 0.158 | 0.195 |
| LCUI Quadratic | 0.142 | 0.116 | 0.143 | 0.175 |
| LCUI 2$^{nd}$ Order Approx. (Linear) | 0.212 | 0.172 | 0.157 | 0.198 |
| LHET Linear | 0.160 | 0.160 | 0.157 | 0.154 |
| LHET Quadratic | 0.151 | 0.148 | 0.158 | 0.156 |
| LHET 2$^{nd}$ Order Approx (Linear) | 0.145 | 0.146 | 0.144 | 0.142 |
| EHET Linear | 0.304 | 0.284 | 0.284 | 0.304 |
| EHET Quadratic | 0.291 | 0.271 | 0.273 | 0.295 |
| EHET 2$^{nd}$ Order Approx. (Linear) | 0.301 | 0.282 | 0.275 | 0.295 |

In terms of the different ranges of capacity utilisation, the results for LHET are very similar to the original ones. This suggests that the results are relatively stable. In contrast however, the removal of cells with capacity utilisation below 30% or below 30% and above 75% generally produce a noticeably reduced adjusted R-squared result for LCUI. In particular, the results following the removal of cells with capacity utilisation below 30% show that for this measure these low levels make an important contribution to its accuracy. However, the removal of just cells with capacity utilisation above 75% produces a substantially improved result for the Linear and Quadratic functional form but a worse position for the Second Order Approximation (linear) form. This shows the measure performs less well at high levels of capacity utilisation for two of the functional forms. These results suggest that, unlike LHET, the LCUI measure performs better at certain levels of utilisation than others.

The EHET measures results are worse for the three functional forms when cells below 30% are removed and when cells below 30% and above 75% LCUI utilisation are removed. These two sets of results are very similar. Finally, the results following just those cells with a utilisation above 75% are very similar to the original results. The sensitivity tests for the EHET capacity utilisation measure therefore suggests that its accuracy increases following the inclusion of cells with low levels of capacity utilisation.

It can be seen that the EHET capacity utilisation measure with a Linear Functional Form continues to have the highest adjusted R-squared. Given that EHET has a much higher adjusted R-squared result no matter the option shown in Table 7.23 this remains the preferred option.

Despite being contrary to expectations, the results of the analysis carried out for this thesis show that the Linear functional form produces the best result for the area capacity utilisation measures. One possibility for this unexpected finding is that the Data Set with just 64 observations is simply too small to produce conclusive results.

## 7.5 β Values for the Area Capacity Utilisation Measures

This section presents the calculated β values for the three area capacity utilisation measures using the Linear functional form with a one-way, fixed effects approach. These are given in Table 7.24.

**Table 7.24** Calculated β's for the Area Variables (Linear Functional Form and a One-Way 'Fixed' Effects approach).

| Variable | Scope | Calculated β |
|---|---|---|
| LCUI | Part ECML | 0.0129 |
| LHET | Part ECML | 0.0138 |
| EHET | Part ECML | 0.0203 |

It can be seen that the calculated β is highest for the preferred area capacity utilisation variable (EHET). The other two area variables (LHET and LCUI) have very similar β values. Once again the individual '*A*' coefficients also have an important impact on the modelled levels of CRRD. The full regression results are presented in Appendix C.

## 7.6 Overall Summary of Regression Results

For the sectional data set, the Exponential fixed effects one-way model is the preferred approach to exploring the relationship between capacity utilisation and reactionary delay. Although, the Second Order Approximation (logarithmic) functional form produces the 'best' fit in the majority of cases, as previously discussed this possibly reflects the nature of the data set. As previously noted, the choice of the Exponential functional form is both intuitive and consistent with previous work on the subject.

For the area data set, the Linear fixed effects one-way model provides the 'best' approach. As discussed previously though a Linear functional form is contrary to expectations and is not consistent with previous work.

In terms of the sectional variables, the HET based variables provide the best results (with XHET being the most preferred). Although still significant, the results for the CUI based variables are surprisingly not too dissimilar to those for Intensity. This demonstrates that although CUI is a useful measure of the volume of capacity utilisation, how capacity is used in a timetable is a more effective determinant of performance (as measured by the level of reactionary delay).

As noted, the sectional capacity variables produced more than reasonable adjusted R-squared values. The inclusion of 'other' variables in an attempt to produce a better 'fit' of the curve; although demonstrating that a number of factors other than capacity utilisation in the section in question were significant were not felt to add to the explanatory power of the capacity utilisation measures significantly enough to warrant their inclusion. The ‘complexity of the timetable’ was however found to be an important determinant of reactionary delay even when capacity utilisation was not taken into account.

One possible reason why a better ‘goodness of fit’ from these additional variables was not achieved is the fact that permanent timetable data (adjusted to remove non-running or rarely running trains) was used. As described in Chapter Six, greater accuracy would have been achieved by using the services that actually ran on the day. However, given the number of days in the data set only the representative day approach is practical and as noted previously this has its own disadvantages.

In terms of the area capacity variables, EHET provides the best results followed by LCUI and then LHET. This suggests that for wider areas the minimum gaps between trains wherever they occur provide a better indication of overall levels of reactionary delay than capacity utilisation at the primary constraint. A possible explanation of why LCUI performs better than LHET has been provided in section 7.4.5.

## 7.7 Choice Between Sectional and Area Explanatory Variables

Due to the difference in the size of the data sets it is obviously not possible to make a direct comparison between the area and sectional capacity variables. However, the substantially higher adjusted R-squared scores for the sectional capacity variables, despite the fact that with the sectional data there are six times as many data points to ‘fit’ the regression curve; does suggest that analysis at the sectional level provides a better result than an at the area level. Furthermore, the adoption of a Linear functional form for the area analysis is as discussed neither consistent with previous work or intuitive.

The conclusion that a sectional approach is preferable to an area one is intuitive as the measured capacity utilisation matches the specific location that the reactionary delay has been recorded for. This means that reactionary delay is more affected by capacity utilisation on the specific section in question rather than by a ‘close-by’ constraint. The sectional capacity variable XHET using an Exponential functional form with a one-way, fixed effects approach is therefore preferred of all the options considered in this analysis.

## 7.8 Reasons for the Better Performance of the HET Based Measures

### 7.8.1 Overview

The HET based measures are therefore more effective predictors of reactionary delay in the sample network than the CUI measures. This is intuitive, since as described in Chapter Three previous research has concluded that it is the gaps between trains that determine the resulting level of reactionary delay. In Chapter Three, the two elements of heterogeneity were also discussed. These are firstly a mixture of services with different characteristics and secondly the 'bunching' of similar services. Although, CUI accounts for differences in traffic speed it does not recognise the impact of similar traffic 'bunching'. As demonstrated in the examples given in Chapter Three, CUI effectively assumes that traffic with the same characteristics is evenly spaced.

In this section, examples of 'bunching' in the data set are given. The reasons why 'bunching' might occur generally are then discussed.

### 7.8.2 Examples of 'Bunching' in the Sectional Data Set

Three examples are presented in Table 7.25 from the sectional data set which demonstrates the presence of 'bunching' and its association with a higher level of reactionary delay per train mile. These are taken from the Welwyn Fast lines. This is due to the higher volume of traffic than the Welwyn Slow lines and the fact that in the Welwyn area there is less freight traffic and therefore any heterogeneity is more likely to be due to the 'bunching' of traffic with the same characteristics.

It can be seen that two time periods are presented for each of the three geographic sections presented in the Table. They have the same number of trains in each time period. The identical (or almost identical) levels of CUI between the two time periods in each case show that there is little variance in the volume of capacity used. However, it can be seen that in each case one time band has a higher level of reactionary delay per train mile (RDTM) than the other. In each case this is matched by a higher level of HET. Since, the volume of capacity is the same it can be concluded that this is due to an increased level of train 'bunching'.

**Table 7.25** Examples from the Sectional Data Set which Illustrate the Greater Effectiveness of HET.

| Time Period | Section | Line | RDTM | No. Trains | OCUI % | OHET % |
|---|---|---|---|---|---|---|
| 0900-1000 | Hitchin to Stevenage | UF | 0.87 | 8 | 45.0 | 50.4 |
| 1000-1100 | Hitchin to Stevenage | UF | 6.17 | 8 | 45.0 | 59.5 |
| 0900-1000 | Hitchin to Sandy | DF | 0.29 | 4 | 26.7 | 41.2 |
| 1400-1500 | Hitchin to Sandy | DF | 0.00 | 4 | 26.7 | 28.0 |
| 1600-1700 | Welwyn Viaduct | DF | 2.87 | 13 | 69.2 | 70.4 |
| 1900-2000 | Welwyn Viaduct | DF | 4.64 | 13 | 70.0 | 84.0 |

### 7.8.3 Possible Reasons for Timetable 'Bunching'

From a commercial and service provision view point it makes sense for passenger services to be evenly spaced in the timetable. The analysis described previously in this chapter also clearly supports the conclusions of other research, described in Chapter Three, that 'even spacing' is more effective at reducing the level of reactionary delay.

There appears to be a number of possible reasons why timetable 'bunching' occurs despite the disadvantages. Firstly, a mixture of trains in the timetable will obviously lead to an uneven spacing of trains. As noted, both the CUI and HET based measures account for this type of heterogeneity. This can be divided into the impact of passenger trains having different calling patterns en-route and the overall influence of mixed traffic.

As outlined in Chapter Six, there is a general feeling that East Coast trains' non-standard calling patterns on the ECML route in the timetable used for the analysis contributed to inefficient capacity utilisation. As explained this was one of the reasons for choosing the route and timetable in question.

Table 7.26 shows the stopping patterns at the three stations for the Up Newark portion of the sectional data set between 1000 and 1200 hours. The table also shows the origin for each of the nine services in the time period. It can be seen that between the three train operators there are six different origins. In the case of the principal Train Operator on the route there are four different origins (Glasgow, Edinburgh, Newcastle and Leeds). It can be seen that the three stations each have a different number of services calling at them and the stops themselves are not evenly spaced. In particular, Retford has only two stopping services in the two hour period which are by successive trains. The fact that they are by different train operators raises the possibility that, despite the comments about competition that will be made later in this section, there is still some competition between rival operators once their access rights have been granted.

The table also shows the total journey time from Retford to Grantham for each of the services. It can be seen that there is a 7 minute difference between the fastest (a non-stop train) and the slowest (a train stopping twice). This will inevitably lead to timetable 'bunching' even in other sections where trains have identical characteristics as any even-spacing on the route will be potentially be disrupted.

**Table 7.26** Stopping patterns on the ECML (Retford to Grantham) in the Sample Timetable from 1000 to 1200 Hours.

| **Train**<br>**Origin** | **1E03**<br>**EC**<br>**Edin** | **1A19**<br>**EC**<br>**Lds** | **1A20**<br>**EC**<br>**Nwc** | **1A21**<br>**EC**<br>**Lds** | **1E05**<br>**EC**<br>**Glas** | **1A93**<br>**HT**<br>**Hull** | **1A61**<br>**GC**<br>**Sund** | **1A22**<br>**EC**<br>**Lds** | **1A23**<br>**EC**<br>**Lds** |
|---|---|---|---|---|---|---|---|---|---|
| Retford | - | - | - | - | Stop | Stop | - | - | - |
| Newark | - | Stop | - | Stop | - | - | - | Stop | - |
| Gthm. | Stop | - | - | Stop | - | Stop | - | Stop | Stop |
| Journey Time | 21 mins | 23 mins | 18 mins | 25 mins | 22 mins | 23 mins | 19.5 mins | 25 mins | 20.5 mins |

(Please see Abbreviations section for a key to those used).

The 'concertina' effect on timetable spacing can also be seen when a mixture of traffic is considered. It is important to appreciate that the presence of mixed traffic as seen on the ECML considerably increases the difficulty of maintaining an even-spacing between trains. Table 7.27 shows an example of this 'bunching' effect by showing the gaps between and after a freight service (4L78) which is 'sandwiched' between two passenger services. The example has also been chosen as in this instance 4L78 is planned to depart Claypole Loop immediately behind a passenger service. The Table also shows the reactionary delay associated with 4L78 and the following passenger train. This is the total observed reactionary delay for the timetable period in question. This demonstrates that the size of the arrival gap at the end of the section is not necessarily the most important gap in determining the level of reactionary delay (and this illustrates why AHET was found to be the least effective of the HET based measures).

The freight train 4L78 is planned to depart Claypole loop the minimum time behind the preceding passenger train (which is less than the general planning headway of 4 minutes). However, by Grantham it is planned to be 10 minutes behind the faster passenger train. The next passenger train (1A28) which at Claypole loop is 13 minutes behind 4L78 is only 4.5 minutes behind by the time Grantham is reached (the distance between Claypole loop and Grantham is only 10 miles). The example therefore not only shows how capacity is quickly used up when trains with different speeds operate on the same line. It also demonstrates how the gaps between trains quickly decrease due to the impact of speed differentials.

**Table 7.27** Illustration of the 'Concertina' Effect on Spacing Caused by Mixing Freight and Passenger Traffic.

| Location | Gap between 4L78 and previous train | Gap between 1A28 and previous train (4L78) |
|---|---|---|
| Claypole Loop | Departs 2 minutes after. | 13 minutes |
| Grantham | 10 minutes | 4.5 minutes |
| Total Observed Reactionary Delay Incurred by Train in Question (within Section) during the Timetable Period. | 47 minutes | 23 minutes |

As noted, the table also shows the observed amount of reactionary delay incurred by each train. It can be seen that 4L78, which starts the section closely behind the train in-front, incurs twice as much reactionary delay as 1A28, which almost catches 4L78 by the end of the section. Clearly, this is a result of 4L78 being 'held' in Claypole loop to allow late running but faster passenger trains to pass thereby preventing even greater levels of reactionary delay.

However as discussed previously, the ‘bunching’ of traffic with identical or very similar characteristics also occurs in the data set and this helps explain the superiority of the HET based measures to the CUI ones. There appears to be a number of possible reasons for this ‘bunching’.

These are:-

- Competition
- The ‘coming together’ of services with different origins and destinations on the same section.
- The ‘need’ to plan timetables around infrastructure constraints.
- Timetable Evolution versus Timetable Revolution
- The impact of restrictive track access agreements.
- Reservation of ‘spare’ paths.

There is also Watson’s suggestion (2008), outlined in Chapter Three, that although ‘even spacing’ might be preferable on links; at junctions overall delays might be reduced by timing trains close together to ‘clear’ the node as quickly as possible. It is also necessary to consider this possibility as it runs counter to the conclusions so far reached in this thesis. This will therefore be examined in a subsequent section.

#### 7.8.3.1 Competition

The introduction of competition was one of the stated aims of the privatisation of Britain’s railways. However, Preston (1999, p18) raises the issue of whether competing services “make the best use of limited capacity”. There is also evidence that in the early days of privatisation there was some ‘predatory’ behaviour as one train operator sought to gain a commercial advantage over another. Wolmar (1996) gives the example of a Train Operator planning a service directly in-front of an existing half-hourly service.

Competition is however unlikely  to be an important factor in the ‘bunching’ of services on a route. This is due to the protection of existing franchised passenger services through legislation which has applied in some form since privatisation. For example,  in the ‘Final Conclusions Report’ on the Moderation of Competition (ORR, 2004b, p17) the then Regulator concluded that “whilst on-rail competition between operators can bring benefits to passengers, there will in practice be limited scope for such competition to develop in the foreseeable future”. In order to counter ‘predatory’ behaviour this legislation meant that only those new services which could be demonstrated to serve new markets and generate new trips would be permitted. Those services which were purely intended to abstract revenue from existing services would not be allowed. Interestingly, in making his conclusions the Rail Regulator made the comment that congested routes were more likely to attract competition and thus further increase congestion, as these would be associated with higher passenger numbers and thus higher potential revenues (ORR, 2004b, p17).

#### 7.8.3.2 The ‘Coming Together’ of Services

The effect of the ‘coming together’ of services with different origins and destinations on the same section is that the timetable ceases to be self-contained. This means that timetabling decisions on the route in question may have to take into account the interaction with services that may themselves have no contact with the route and may be many miles away from it. This is likely in Britain due to the highly inter-connected nature of its rail network. Table 7.26 showed that nine trains in a two-hour period had between them six different origins. Timetable Complexity and Delay Propagation have both been discussed as key issues for congested rail networks earlier in this thesis.

#### 7.8.3.3 Planning ‘Around’ Constraints

There are three key infrastructure constraints in the data set used in this analysis, namely Welwyn Viaduct, Hitchin Cambridge Junction and Newark Flat Crossing. In the case of Welwyn Viaduct the need to coordinate the timings of traffic on the adjacent Fast and Slow lines so that they successfully merge on the viaduct itself will inevitably affect ‘timetable’ spacing. This is illustrated in Table 7.28. This shows the Up Slow timetabled gaps on the section between Stevenage and Woolmer Junction for the 1700 to 1800 time period. It also shows the use of ‘pathing’ time on the section and the gaps to the previous train on the Viaduct itself (i.e. once the Fast and Slow line traffic has combined).

The table shows that at Stevenage the six trains are irregularly spaced. The spacing between these trains is changed through the application of ‘pathing’ time to four of them. In the case of three trains this is clearly to facilitate integration with the Fast line traffic on Welwyn Viaduct. For example, 2C76 has two minutes time added to its journey to make it the minimum three minutes behind the preceding ‘Fast’ train on the Viaduct. Looking at spacing on the Viaduct itself shows that four of the six trains are now timetabled the minimum distance behind the train itself.

**Table 7.28** The Impact on Timetable Spacing of Merging Slow Line and Fast Line Traffic at Welwyn Viaduct (1700-1800 Time Period). The Timetabled Gap for Slow Line Traffic to the Previous Train in Minutes.

| **Location / (Allowance)** | **1C92 Gap** | **2P75 Gap** | **3P25 Gap** | **1P75 Gap** | **2C76 Gap** | **3C26 Gap** |
|---|---|---|---|---|---|---|
| Stevenage | 13.5 minutes | 4 minutes | 14.5 minutes | 9.5 minutes | 6.0 minutes | 11.0 minutes |
| (Pathing Time)* | (0.5) minutes | (0.5) minutes | - | - | (2.0) minutes | (2.5) minutes |
| Woolmer Junction | 11.5 minutes | 5.5 minutes | 12.0 minutes | 10.0 minutes | 10.0 minutes | 9.0 minutes |
| Welwyn Viaduct FL & SL Traffic | 3 minutes | 3 minutes | 3 minutes | 9 minutes | 3 minutes | 5.5 minutes |

* Including Pathing Time in the table demonstrates the level of adjustment to Slow line schedules between Stevenage and Welwyn Viaduct.

However, 3C26 in the Table has the highest level of pathing time in the approach to Welwyn Viaduct (2.5 minutes) despite not needing any to achieve the minimum three minute gap to the previous train. Analysis of the timetable shows that 3C26 is timed on the Fast line until Potters Bar where it crosses onto the Slow Line behind a stopping passenger train. As discussed elsewhere in this thesis, validation of the timetable for an overall route is likely to produce a different outcome to one validated for individual constraints or sections.

In terms of the two junctions, one common approach to minimising capacity utilisation is the use of ‘parallel’ moves (i.e. the coordination of crossing moves in opposite directions so that they are timed to take place at the same time). The December 2008 to May 2009 timetable for both Hitchin Junction and Newark Flat Crossing contain many examples of parallel moves. However, although the strategy clearly saves capacity at important infrastructure constraints it does mean that the general timetable structure is further tied to a specific location.

It is also worth noting that Hitchin Junction and Newark Flat Crossing have very different impacts on the ECML timetable. At Hitchin Cambridge Junction, traffic joins and leaves the main line. There is therefore the impact of trains with different origins and destinations on the structure of the ECML timetable. There is also the added complexity that 'Down' traffic heading towards Cambridge crosses the 'Up' lines at the junction. This need to co-ordinate both directions will inevitably have an impact on timetable spacing on individual lines. Newark Flat Crossing is less complex since movements across the junction are purely crossing moves i.e. traffic from Newark or Lincoln does not join or leave the ECML at this point.

As explained in Chapter Six, there are also other infrastructure constraints in the sample area and these will also have an impact on timetable spacing. A prime example is the spacing between freight loops. This dictates how freight traffic is timetabled on the ECML and as demonstrated by Table 7.27 their use has an impact on timetable bunching. All these infrastructure factors add to the complexity of producing a timetable that delivers its objectives and helps explain the difficulty and impracticality of expecting services to be evenly spaced.

Finally, simply the need to time trains over long distances will have an impact on the specific characteristics of the timetable at a particular location. As discussed on page 134 of this thesis, taking into account the restrictions imposed by both the Newark and Welwyn area; means that long-distance services in particular will have their entry times at each location jointly determined.

#### 7.8.3.4 Timetable Evolution Versus Revolution

Creating a timetable completely from scratch, in other words a timetable revolution, clearly raises the possibility of spacing services as evenly as factors such as infrastructure and 'traffic mix heterogeneity' allow. This will be considerably harder where new services are added into suitable gaps between existing services, in other words timetable evolution. The latter type of timetable will however be clearly easier and quicker to produce. It will also require less agreement with existing operators to introduce new services. The ECML timetable used in the data set is clearly of the evolution rather than revolution type. One example of this is the fact that the services of the two Open Access Operators (Hull Trains and Grand Central) appear to be 'fitted around' those of the existing Franchised Passenger Operators.

This is illustrated by Figure 7.6 which shows the gaps between services on Welwyn Viaduct in the Up Direction in the 1000-1100 time period where both Grand Central and Hull Trains have timetabled services. The gaps are shown in the order that the trains appear in the timetable. The two Open Access services are the third and eighth trains in the sequence. The figure shows that both Open Access trains are accommodated in two of the biggest gaps in the timetable (a gap of 9 minutes and 11.5 minutes respectively)[42] . However, they themselves are timed the minimum planning headway behind the train in-front demonstrating that they are both subject to timetable 'bunching'.

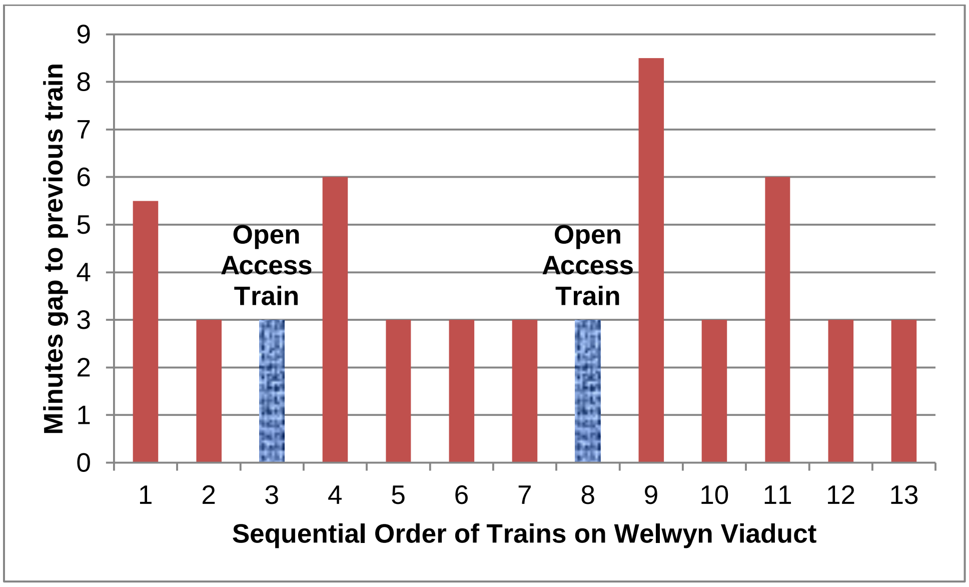


**Figure 7.6** Timetable Spacing on Welwyn Viaduct in the Up Direction (1000-1100 Time Period)

### 7.8.3.5 Restrictive Track Access Agreements

One feature emerging from the privatisation of Britain's rail network was the creation of a large number of legal agreements between the newly separated operational parts. In their 2004 Guide to the Model Passenger Track Access Contract, the ORR (2004a, p8) noted that "it is through the track access contract that an operator is granted access to the network and hence the capacity of the rail network is shared out".

[42] This is calculated by adding together the gap for the Open Access Train itself and the one behind it together (i.e. Train 3 + Train 4 and also Train 8 + Train 9).

One noticeable feature of the track access contracts is the significant number of elements of a timetable that are listed and therefore protected. Schedule 5 specifies the services that each train operator is entitled to. There are two types of right: Firm Rights and Contingent Rights.

Broadly speaking, Firm Rights are only subject to any contractual right that Network Rail has to flex trains and the provisions of the agreed timetable planning rules and Network Code.

Less protected are any Contingent Rights as these are also subject to other factors such as the firm rights of other operators. ”Contingent rights may not always be satisfied, and space in the working timetable to meet all operator’s firm rights is always allocated before any space for contingent rights” (ORR, 2004a, p11).

From a commercial perspective it is therefore in a train operator’s interest to have as much of their services specified as Firm Rights as possible since this pretty much guarantees their delivery. Many aspects of a service can be given Firm Rights (e.g. Departure Times, Arrival Times, Journey Times and Calling Patterns). Therefore, although the ORR has stated that “it has never been the Regulator’s intention to make the model contract a straitjacket” (ORR, 2004a, p2) there is clearly a risk that the benefit to operators of securing firm rights coupled with the pressure on Network Rail to fulfil all of them may result in the inefficient use of capacity.

The ORR has made some attempts to reduce the rigidity of Track Access Agreements through for example the formation of an Industry Working Group. However, this met opposition from Operators concerned about risk to their revenue. In summing up the output the ORR stated that they “still believe there is scope for simplifying the expression of access rights. This remains an important issue that needs to be addressed as the network becomes increasingly congested and given the move towards longer-term franchises. However, given current views we think it would not be appropriate to pursue this work-stream at present” (ORR, 2010, p13). Instead, their intention was to continue to consider the level of detail requested in new track access contracts on a case by case basis.

#### 7.8.3.6 The Reservation of ‘Spare’ Paths

Finally, capacity may be kept reserved in a timetable in the form of ‘unused’ paths. This means additional traffic can be incorporated as demand arises avoiding the need for major timetable revisions. Such an approach is particularly suitable for freight traffic; due to the difficulty of predicting the long-term demand for paths. The incidence of unused or rarely used freight paths in the sample ECML timetable has already been discussed with the paths themselves presented in Appendix A.

The approach also allows the efficient utilisation of capacity from the perspective of actual occupation. This is because the number of gaps in a timetable too small to accommodate an additional service is kept to a minimum. This aspect of capacity utilisation is also clearly complimented by the CUI based measure of capacity utilisation with its ‘compression’ methodology; which indicates how ‘full’ a particular timetable is and the potential for additional paths. However, the approach is clearly contrary to a policy of even-spacing which as demonstrated earlier in the chapter is supported by the HET based measures of capacity utilisation.

## 7.9 The Preference for Evenly Spaced Timetables

The conclusion that XHET is the most effective of the capacity utilisation measures considered suggests that the even-spacing of traffic on links and at junctions will minimise levels of reactionary delay. This conclusion matches the majority of the research referred to in Chapter Three (e.g. Carey, 1999).

However, Watson (2008) advanced an alternative view that although even-spacing might be the best approach on less complicated networks; the ‘flighting’ of traffic will be more effective at reducing delays on more complex layouts. At junctions this suggests that increasing the size of the gap between ‘conflicting’ moves at junctions at the expense of decreasing the gaps between trains in the same direction (i.e. ‘flighting’ them) is preferable to an even-spacing for all trains. The latter is the assumption made during the calculation of the XHET capacity utilisation measure.

This was investigated using data from the sample network.  The impact of crossing moves at Hitchin Junction on the levels of reactionary delay data for the adjacent approach links was considered. For the analysis the Sandy to Hitchin Up Fast line was chosen. The Up Fast is crossed by Down traffic heading onto the Cambridge Branch. However, in the majority of cases the corresponding Up traffic from the Branch joins the Up Slow. Up Fast capacity utilisation is therefore generally effected by two types of flow (i.e. the through ‘Up’ flow between Sandy and Hitchin and ‘Down’ crossing moves at Hitchin Junction).

Table 7.29 examines the relationship between individual trains in the 0800-0900 time period. The table shows that the highest amount of reactionary delay is associated with a through train with the smallest ‘buffer’ following another through train. Indeed, it can also be seen that the bigger the buffer behind a through train the smaller the observed amount of reactionary delay. The picture where the previous train is a crossing ‘conflicting’ move is however much more mixed. Although the largest amount of reactionary delay in this group is associated with the smallest ‘buffer’, a train with a large buffer (1A08) also has a substantial amount.

Table 7.29 therefore suggests that the relationship between small buffers and increased reactionary delay applies to all types of traffic move (which supports the conclusion that XHET is an effective measure of capacity utilisation). However, the data set clearly includes some exceptions to this rule which underlines the impact of ‘other’ factors on the level of observed reactionary delay.

**Table 7.29** Comparison Between Buffer Times and Associated Levels of Reactionary Delay on the Sandy to Hitchin ‘Up Fast’ Section (0800-0900 hours).

| Period (Start Time) | Train Headcode | ‘Buffer’ in minutes | Previous Type of Train | Reactionary Delay (minutes) | Reactionary Delay as % of Hourly Total |
|---|---|---|---|---|---|
| 0800 | 1P54 | 7.5 | Through | 7 | 4.3 |
| 0800 | 1A05 | 0.5 | Through | 62 | 38.0 |
| 0800 | 1A06 | 1.0 | Crossing | 3 | 1.8 |
| 0800 | 1A07 | 0.5 | Crossing | 3 | 1.8 |
| 0800 | 1A08 | 5.0 | Crossing | 24 | 14.7 |
| 0800 | 1P55 | 1.0 | Through | 11 | 6.7 |
| 0800 | 1A09 | 0.5 | Crossing | 31 | 19.0 |
| 0800 | 1A91 | 0.5 | Crossing | 22 | 13.5 |

It is however worth paying further consideration to the relative merits of even-spacing and ‘flighting’ generally on a network. Whilst ‘even-spacing’ has the advantage that all trains have the same size ‘buffer’; ‘flighted’ trains have the advantage that there is one (or several) larger ‘fire-breaks’. Since, a primary incident can theoretically occur at any time; the starting point of any reactionary delay can therefore be at any point in a sequence of trains. This means that ‘even-spacing’ will be more effective, since every train has a buffer preceding it; providing that the reactionary delay is small enough to be absorbed without serious propagation.

As discussed, in Chapter Six the average size of delay in the Data Set is relatively small (an overall mean  of approximately four minutes per occasion). This means that for lines with three minute planning headways, a traffic intensity of 50% or under (i.e. 10 out of a possible 20 trains an hour) and perfect ‘even-spacing’ a four minute initial delay would result in the following train only ‘suffering’ 1 minutes reactionary delay. Even for a traffic intensity of 75% (i.e. 15 trains per hour) with ‘perfect-even’ spacing’, the ‘buffer’ of 1 minute per train would mean that four minutes initial delay would be completely absorbed having caused the following three trains a total of six minutes reactionary delay. In fact the lower  median than mean suggests that the data is right-skewed which implies that in many cases no reactionary delay will be generated at all following the initial delay.

‘Flighting’ will be of greater benefit where the volume of traffic and or the size of delay is such that even-spacing would lead to individual ‘buffers’ for trains being insufficient to prevent serious propagation of delay. There is of course the issue of where in a sequence of trains the ‘firebreak’ is placed. As the results of the analysis show, the concept of the gaps before ‘vulnerable’ trains having a greater influence on the level of delay is an important one. Placing a ‘firebreak’ so that trains which have an inter-connection with other parts of the network are most protected could therefore be the most advantageous strategy. However, the generally small size of delays in the sample data set suggests though that except for the very highest volumes of traffic, an evenly spaced timetable would tend to be more effective at reducing the overall level of delays than a ‘flighting’ strategy. It is also worth noting that the HET based measures could take the existence of a ‘flighting’ strategy into account by weighting the ‘firebreak’ gaps. Such an exercise is however beyond the scope of this thesis.

It is clear though that either planned ‘spacing’ strategy will be more effective at reducing reactionary delay than the generally irregular pattern of spacing seen in the sample data set. Given the expectation that a primary incident can occur at any point during a sequence of trains, there is therefore an equal chance that any train in the sequence will incur primary delay. The existence of ‘bunching’ in a timetable means that overall, traffic will tend to incur more delay than would occur if the timetable was evenly spaced. This is the reason why the HET based measures, which account for timetable ‘bunching’, are more effective than the CUI based measures that do not.

## 7.10 The Influence of Constraints on Capacity Utilisation and Performance

As described earlier in this chapter, although the sectional capacity measures with their greater detail are considered better predictors of reactionary delay, both the area explanatory variables which measure the capacity utilisation of the primary constraint (LHET and LCUI) were still found to be significant. It is therefore worth exploring the influence that the main infrastructure constraints in the sample network have on both capacity utilisation and reactionary delay.

Figures 7.7 and 7.8 compare the calculated capacity utilisation using the XHET measure and the recorded RDTM for geographic sections associated with the two parts of the ECML included in the data set.

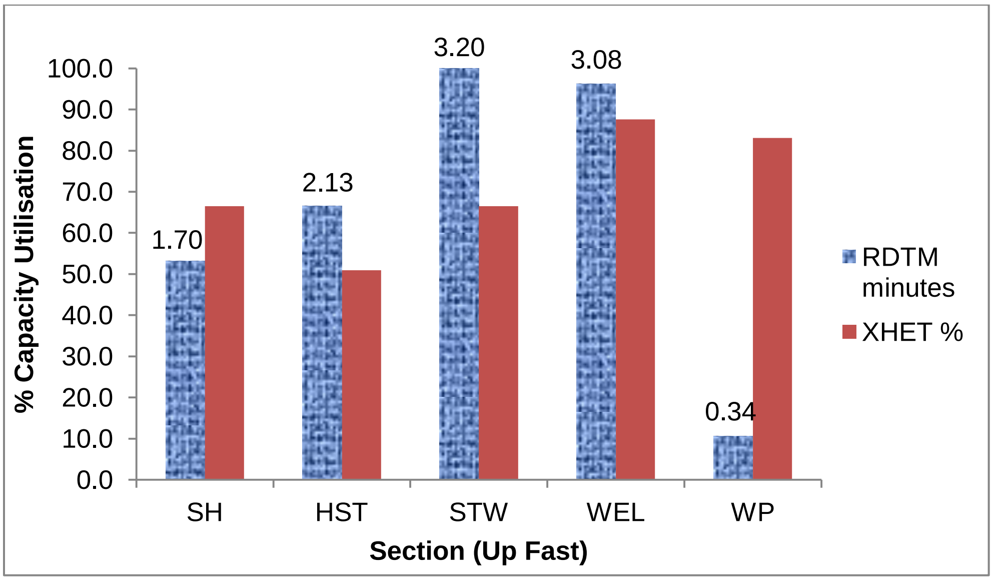
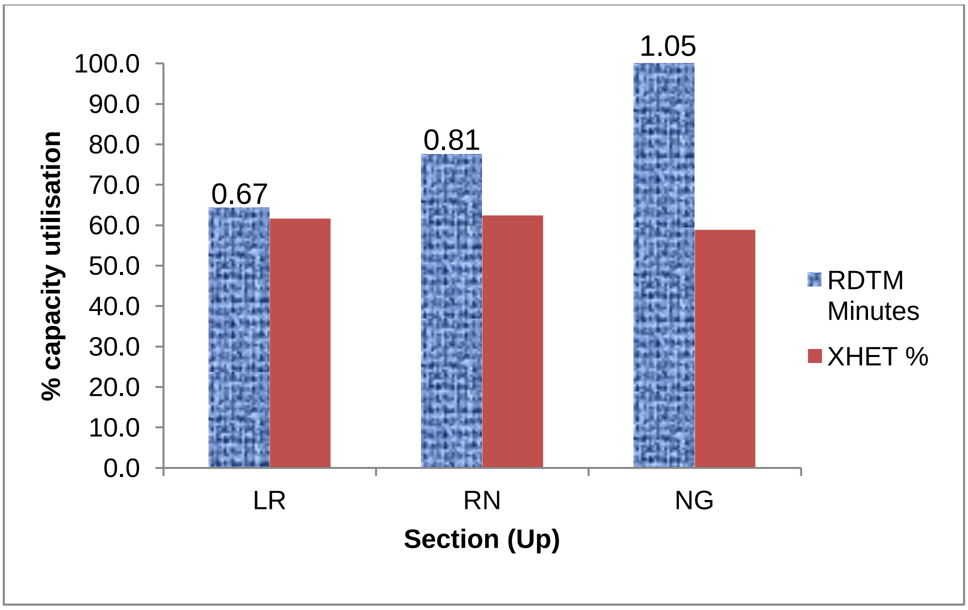


**Figure 7.7** Comparison of % XHET Capacity Utilisation and Minutes RDTM for Welwyn ‘Up Fast’ (0800-0900)

Figure 7.7 covers the five ‘Up Fast’ sections associated with the Welwyn area which includes the Hitchin Junction and Welwyn Viaduct infrastructure constraints. These are Sandy to Hitchin (SH); Hitchin to Stevenage (HST), Stevenage to Woolmer Green Junction (STW); Welwyn Viaduct (WEL) and Welwyn Garden to Potters Bar (WP). A number of interesting observations can be made. Firstly, Welwyn Viaduct itself has the highest capacity utilisation but only the second highest level of reactionary delay. The highest level of reactionary delay is observed on the section in advance of Welwyn Viaduct (Stevenage to Woolmer Green Junction). Both sections approaching infrastructure constraints (Sandy to Hitchin and Stevenage to Woolmer Green Junction) have the same XHET %, however the latter has a much higher level of reactionary delay per train mile.

Furthermore, although the section in advance of Hitchin Junction has a higher calculated level of capacity utilisation it has a lower level of reactionary delay than the section after it. The steady rise in recorded RDTM up to Welwyn Viaduct suggests a combined effect of the two infrastructure constraints on timetable performance. Finally, although Welwyn to Potters Bar has a high degree of capacity utilisation the observed level of reactionary delay is very low. This shows the variation in reactionary delay that can be seen in adjacent geographic sections with similar levels of capacity utilisation. This underlines the value of the section coefficients (i.e. ‘A’) in the specification, as described earlier in this thesis.



.

**Figure 7.8** Comparison of % XHET Capacity Utilisation and Minutes RDTM for Newark ‘Up’ (0900-1000).

Figure 7.8 shows the three sections that make up the Newark area in the Up direction. These are Loversall to Retford (LR); Retford to Newark (RN) and Newark to Grantham (NG). The impact of the junction moves at Newark Flat

Crossing is taken into account in the capacity utilisation for the Retford to Newark section. The diagram compares the calculated XHET percentages for the three sections with the observed levels of reactionary delay. It can be seen that calculated capacity utilisation is very similar, despite only the middle section having a significant infrastructure constraint. There is however a steady rise in the observed levels of reactionary delay. The reason for both these observations may be the fact that capacity utilisation and also timetable robustness for this part of the network is influenced by a significant mix in traffic type. Figure 7.8 also suggests a cumulative effect on levels of reactionary delay from the three geographic sections.

Therefore, although infrastructure constraints do have an influence on the level of reactionary delay for a network; the nature of individual geographic sections also plays a very important role. This reinforces the conclusion given in Section 7.7 that analysis at the sectional level is more effective than at the area level.

## 7.11 The Representativeness of the Conclusions from the Data Set

As noted previously, the sample data set consists of two small sections of one of Britain's primary rail routes. An important final question is how applicable are the conclusions discussed in this chapter to other congested parts of the rail network. As described in Chapter Six, the ECML route was chosen for the analysis for its mixture of traffic and infrastructure and the known congestion issues. Although, other routes and even other parts of the ECML will clearly have important differences to the sample network; there will always be an interaction between the volume and type of traffic and the actual infrastructure.

The results clearly show that the HET based measures correctly attribute more reactionary delay to irregular spacing (or 'bunching') than even spacing. They are therefore more effective than the CUI based measures which do not. Therefore, on other routes where irregular spacing is a feature it could be expected that HET would be more effective than CUI. On those routes where there was even-spacing since (as demonstrated in Chapter Three) CUI and HET give the same result, HET would still be an effective measure to use.

There are two potential caveats that have already been mentioned. Firstly, if a route was associated with a high average level of reactionary delay per record or secondly, if there was a very high flow of traffic; the same success of HET might not be seen due to its assumption that even-spacing will reduce reactionary delay. However, even in these cases this could potentially be addressed through the weighting of certain gaps in the HET calculation (as seen in this thesis in the Vulnerable HET measures). For example, ‘firebreak’ gaps could be given a greater weighting than the other gaps in a timetable sequence due to the advantage derived from having them as large as possible. This would suggest that HET would still be more effective than CUI at predicting reactionary delay.

The better performance of XHET is intuitive due to the expectation that the junctions in a network, because of the interaction between different flows, will have an important influence on the observed level of reactionary delay. In Newark Flat Crossing and Hitchin Junction, the sample data set contained two different key junctions with very different characteristics. The XHET approach which was applied in identical fashion to the two junctions however was found to be successful. This suggests that the HET approach is able to cope with junctions of different types and complexities. Once again XHET could potentially be improved by the weighting of certain types of gaps, for example at the moment a ‘through’ gap and a ‘crossing’ gap are given an equal weight. A change to this would however need careful consideration and might vary between locations. As discussed in the previous section, the evidence from the sample data set suggests that contrary to Watson’s (2008) belief, the size of crossing gaps are not a more important factor than the size of through gaps in determining reactionary delay on the approaches to junctions.

The analysis did however exclude stations which are also seen as an important factor in overall levels of reactionary delay. The available platform capacity and any limitations imposed by the track layout which accesses them will be an important determinant on the overall volume of traffic that can be accommodated. As discussed in Chapter Three there has been work by other researchers on applying the CUI approach to station capacity. It should also be possible to apply the HET approach to this as well. In the case of the latter, intuitively the time between a train departing a platform and the next one arriving will determine the likelihood of the latter suffering reactionary delay. One complicating factor though is that at stations there are often different platforms which could be used. The nature of the track layout will also dictate access to individual platforms. Further work is therefore required, which is outside the scope of this thesis, in order to develop a HET based measure which encompasses station capacity utilisation.

The choice of an Exponential functional form is intuitive and consistent with other findings. The preference for a one-way rather than a two-way model is also intuitive, since it seems highly unlikely that other than by changes in capacity utilisation a specific time period should have a direct impact on the level of reactionary delay. As noted this is also consistent with the conclusions of the 2013 recalibration of the Capacity Charge (Arup 2013). The preference for a 'fixed effects' approach rather than 'random effects' approach is also intuitive and consistent. It means that any variation that cannot be explained by the differences in measured capacity utilisation can be attributed to the specific nature of the geographic section in question. These particular findings are supported by (and reinforce) the conclusions of the Capacity Charge recalibration whose scope covered the entire British rail network.

The calculated adjusted R-squared values were found to be more than reasonable particularly given the fact that panel data was used. However, it was disappointing that only modest improvements were gained by the addition of 'other' variables to the regression specifications. It was also surprising that some variables which had been expected to be relevant explanatory factors were found to be insignificant. This is particularly true of those associated with 'network effects'. This is contrary to opinions expressed in Arup's report (2013) on the Capacity Charge recalibration which suggested these could help explain the poor adjusted R-squared values which had been found in that analysis. One explanation is that the use of a one-way fixed effects approach helps account for the influence of factors outside the geographic section in question. Finally, the reason why the adjusted R-squared values are not greater may simply be the day-by-day difference between the timetabled services measured and the actual operated services that produced the observed levels of reactionary delay. The issue of planned services versus operated services will apply to a variable extent through-out the rail network.

The finding that primary infrastructure constraints influence the level of reactionary delay is also likely to be relevant to the rail network as a whole. Except for the simplest networks with the simplest timetable there are always likely to be locations that act as capacity constraints. The complex and interconnected nature of Britain's rail network suggests that there will be few places where this concept does not apply to some extent. However, as found in this analysis the complexity and interconnected nature of the infrastructure itself means that it is likely to be the over-lapping influence of the capacity utilisation in different places which is important. The area analysis demonstrated that EHET which measured the minimum timetable gap anywhere in the network was a more effective indicator of timetable performance than the measures (LCUI and LHET) which measured capacity utilisation at the primary constraints.

Finally, the reasons given for the irregular spacing (or ‘bunching’ of traffic) in Section 7.8.3 are general rather than specific.

## 7.12 Conclusions

This chapter has described the results of the regression analysis and discussed the reasons behind them. The results show that there is a strong relationship between capacity utilisation and timetable performance, as expressed by the observed level of reactionary delay. An Exponential functional form has been found to be the most appropriate and this is both intuitive and consistent with previous research. It is entirely reasonable that as a network becomes increasingly congested reactionary delay will increase at a greater than linear level.

The HET measures which consider the timetabled gaps between trains have been found to be more effective than the CUI based or Intensity variables, which measure the volume of capacity used. XHET, which measures gaps for both through and crossing moves, has been identified as the most effective. The size of the minimum gap is an important indicator of the level of reactionary delay over quite long sections of route (as evidenced by the significance of the EHET variable).

Analysis shows that the ‘bunching’ or ‘irregular’ spacing of traffic has been found to be associated with higher levels of reactionary delay and this explains the better performance of the HET based variables. The reasons for timetable ‘bunching’ has been discussed as has the choice between even-spacing and ‘flighting’ as the most appropriate strategy for minimising delays.

Although, the addition of ‘other’ explanatory variables was not felt to sufficiently improve the strength of the relationship between capacity utilisation and reactionary delay to warrant their inclusion; the complexity of the timetable was found to be an important factor in its own right in determining its performance.

The analysis demonstrates that although the ‘Theory of Constraints’ philosophy has some applicability to the relationship between capacity utilisation and performance; the utilisation on each geographic section appears to have a much bigger influence on the resulting level of reactionary delay. The latter is also more important than ‘network effects’.

However, although the results are more than reasonable it does appear that the strength of the relationship between capacity utilisation and performance is diminished by the difference between timetabled services and those actually operated on the day.

Finally, the transferability of the findings has been discussed with the conclusion that they should apply on a general basis to the network as a whole.

The application of the these findings to the pricing of congested rail networks will be discussed in the next chapter.

# Chapter Eight
# Implications for the Charging of Congested Rail Networks

## 8.1 Introduction

The previous chapter detailed the results of the regression analyses. These show that for the data set in question, a capacity utilisation methodology based on the size of gaps between planned services is superior at predicting levels of reactionary delay, than one based on the volume of capacity used. It has been established that an Exponential functional form is the preferred way of describing the relationship between capacity utilisation and reactionary delay. Furthermore, a one-way fixed effects approach has been identified as the most appropriate one of all the alternatives studied. The relationship between capacity utilisation and performance has been examined at both a sectional and area level. It has been concluded that the sectional approach is the most effective. The likely reasons behind these findings have been discussed in Chapter Seven. Finally, the possible transferability of these findings to other rail networks has been discussed. It has been concluded that the findings are likely to be relevant elsewhere.

This chapter uses the results of the regression analyses to consider alternatives for the charging of congested rail networks. Tariff Equivalents[43] are calculated for both the CUI and HET approaches using a methodology similar to that adopted for the Capacity Charge recalibration (Arup, 2013) and these are then compared. The implications for any differences between the two sets of Tariff Equivalents from a charging perspective are discussed. This chapter also considers what practical alternatives there might be to the methodology previously adopted and compares these results to those already obtained.

## 8.2 Overview of Approach

The calculation of the example Tariff Equivalents described in this section uses the principle of the ‘additional’ train illustrated in Figures 4.3 and 5.1 . As previously discussed this most closely follows the methodology adopted

[43] As discussed in Chapter Five, since the calculations carried out for this Thesis omit some of the information used to produce the Capacity Charge tariffs, notably a monetary value; the figures used here are referred to as ‘Tariff Equivalents’.

for the recalibration of Britain's Capacity Charge. The actual methodology used is outlined in Section 5.4.Tariff Equivalents were calculated using the values for '*A*' and '*β*' for the Exponential fixed effects one-way model. These values are contained in Appendix C.

As previously discussed, the Capacity Charge recalibration used three hour time bands (Arup, 2013). For this thesis Tariff Equivalents were calculated for both hourly and three hourly periods. The three hourly Tariff Equivalents were produced using a simple averaging process.

As discussed in Chapter Five; the Infrastructure Fault values, Schedule 8 payment rates and Lateness:Delay ratio were not taken into account in the calculations described in this chapter. This was due to a lack of data availability. No adjustments have been made to any of the calculations presented here for this reason. Since each capacity utilisation measure has been treated the same the comparison of the results are believed to be still valid. However, one possible issue is the likelihood that the Infrastructure Fault value (or the percentage of reactionary delay that Network Rail is responsible for) differs substantially for different geographic sections. This caveat does therefore need to be borne in mind when considering the results presented later in this chapter. However, any differences will apply equally to each capacity utilisation measure. Finally unlike the Capacity Charge, the Tariff Equivalent was not multiplied by the number of trains in the particular cell. This means that the Tariff Equivalents contained in this chapter are therefore CRRD per train mile (referred to here as RDTM).

Chapter Five described the weighting of the Tariff Equivalents for two different reasons:-

- The combining of Fast and Slow Line Tariff Equivalents.
- The combining of adjacent sectional Tariff Equivalents to produce overall Service Code Tariff Equivalents.

Two different methods were used and these are illustrated in Example Five which is shown in Figure 8.1.

It can be seen that the combining of Fast and Slow Tariff Equivalents is weighted on the basis of train mileage (or effectively the number of trains as the two lines are the same length). This means that Tariff Equivalents for lines with heavier flows of traffic have a greater weighting. However, in contrast the combining of adjacent sections is weighted solely on the basis of mileage. This therefore means that Tariff Equivalents for lines of greater length have a greater weighting. These different approaches reflect the different nature of the relationships with reactionary delay being considered. For parallel Fast and Slow lines different trains are being considered so it is appropriate to weight the Tariff Equivalent by traffic volume. However, for adjacent sections the issue is the length of railway.

**Example Five**

Two Sections A to B

A (10 miles in length); B (7.5 mile)

Both have Fast and Slow Lines.

Calculated Individual Tariff Equivalents:-

A Fast = 0.453 A Slow = 0.991

B Fast = 0.821 B Slow = 0.676

Trains (& Train Miles):-

A Fast = 4 (40) A Slow = 2 (20)

B Fast = 4 (30) B Slow = 2 (15)

**Combining of Fast & Slow Tariff Equivalents (Weighted by Train Mile)**

Section A (0.453 * 40/60) + (0.991 * 20/60) = 0.632

Section B (0.821 * 30/45) + (0.676 * 15/45) = 0.772

**Combining of Adjacent Tariff Equivalents**

Section A + Section B = (0.632 * 10/17.5) + (0.772 * 7.5/17.5) = 0.692

The 'Service Code' Tariff Equivalent for AB is therefore 0.692.

**Figure 8.1** Example Five - Illustration of the Methodology used to Calculate the Sample Tariff Equivalents.

Whilst Fast and Slow lines were combined at this stage there was no consolidation by direction. This recognises the fact that congestion is likely to vary by direction. This is certainly the case for the ECML data set and in particular the Welwyn Area. The combination of Tariff Equivalents by direction would therefore help 'smooth out' any incentive to operate outside the peaks.

Finally, although Tariff Equivalents were calculated for the majority of the capacity utilisation variables described in this thesis; for the sake of brevity only those calculated for OCUI and XHET are presented. OCUI has been chosen as the equivalent to the capacity utilisation measure used in the production of Britain’s Capacity Charge. XHET has been chosen as the most effective predictor of reactionary delay of the explanatory variables considered in this thesis.

## 8.3 Fast and Slow Line Tariff Equivalents

As discussed, the Fast and Slow line Tariff Equivalents were consolidated into a single Tariff Equivalent for each geographical section. In fact many of the Welwyn area Slow lines produce higher Tariff Equivalent than their adjacent Fast lines. This is illustrated in Table 8.1 which shows the calculated Tariff Equivalents for the five sections with the highest RDTMs and their equivalent line pair. The calculated Tariff Equivalents and number of trains per hour are also given.

**Table 8.1** Comparison of Calculated Tariff Equivalents for Five Fast Line / Slow Line Pairs.

| Time Period | Geographic Section | Line | Trains per hour | RDTM | Calculated Tariff Equivalent (XHET) |
|---|---|---|---|---|---|
| 0600-0700 | Stevenage to Hitchin | Slow | 3 | 6.87 | 0.077 |
| 0600-0700 | Stevenage to Hitchin | Fast | 7 | 1.09 | 0.057 |
| 1000-1100 | Hitchin to Stevenage | Slow | 6 | 1.75 | 0.053 |
| 1000-1100 | Hitchin to Stevenage | Fast | 8 | 6.17 | 0.107 |
| 1500-1600 | Stevenage to Woolmer | Slow | 5 | 4.98 | 0.248 |
| 1500-1600 | Stevenage to Woolmer | Fast | 8 | 1.48 | 0.082 |
| 0900-1000 | Woolmer to Stevenage | Slow | 6 | 4.68 | 0.139 |
| 0900-1000 | Woolmer to Stevenage | Fast | 8 | 0.44 | 0.075 |
| 1500-1600 | Woolmer to Stevenage | Slow | 5 | 4.56 | 0.111 |
| 1500-1600 | Woolmer to Stevenage | Fast | 7 | 0.77 | 0.101 |

It can be seen that the higher Slow line Tariff Equivalents reflect the relatively high degree of reactionary delay for relatively low levels of traffic.

As discussed in the previous chapter this imbalance seems to result from the impact of both the two-track Welwyn Viaduct and of traffic joining and leaving the main line at Hitchin Junction on reactionary delay. There is also a possible priority given to fast line traffic since most fast line traffic have the higher priority of a Class 1 head-code compared to the Class 2 head-code associated with the bulk of slow line traffic. If this is true then having separate Fast and Slow line tariffs would unfairly penalise Slow line traffic and increase costs for one type of service compared to another. The most equitable arrangement is to have a single shared tariff.

A further comparison of Fast and Slow line Tariff Equivalents is provided by Figures 8.2 and 8.3. These compare the Fast and Slow line tariffs for the XHET measure for the Welwyn Up direction in the morning peak period (0700 to 1000) and for the Welwyn Down direction in the evening peak period (1600 to 1900) respectively. A line showing the 'consolidated' tariff is also included on each graph.

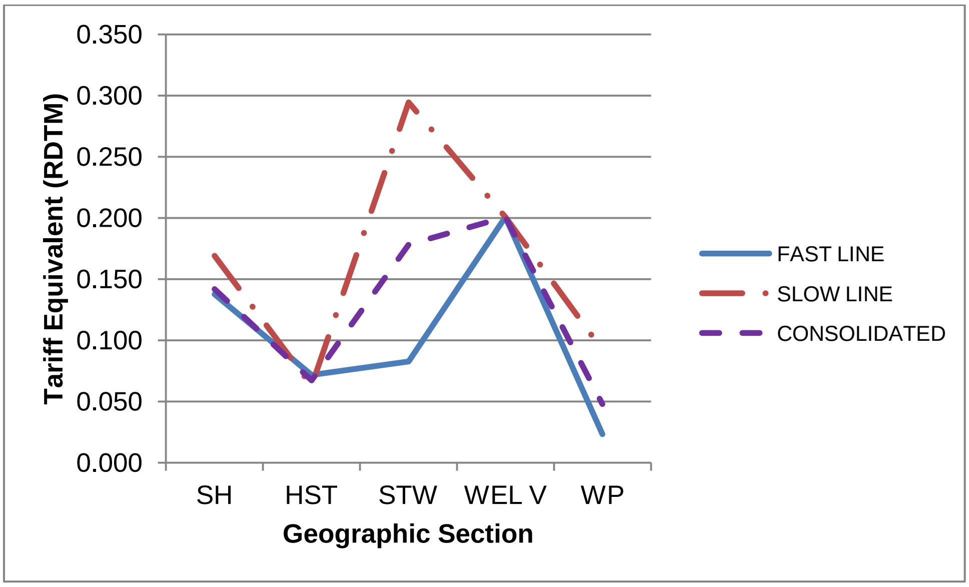


**Figure 8.2** Comparison of Fast and Slow Line Tariff Equivalents for the Welwyn 'Up' Area (0700 to 1000 hours).

The Fast and Slow line Tariff Equivalents in Figure 8.2 appear to be radically different due to the presence of a very high Tariff Equivalent for the Slow line between Stevenage and Woolmer Green Junction (i.e. STW) compared to a low Fast line Tariff Equivalent. This discrepancy reinforces the view that Slow line traffic suffers delay on the approach to Welwyn Viaduct as it waits to merge with the Fast line traffic but can also be due to the interaction with Hertford Loop traffic between Langley Junction and Stevenage. Putting aside this difference though, it can be seen that both Fast and Slow lines have relatively high Tariff Equivalents for the approach to Hitchin Junction (i.e. SH) and for Welwyn Viaduct itself (i.e. WEL V) and low Tariff Equivalents following the constraints (i.e. HST and WP).

Figure 8.3 shows that the Welwyn Down direction has a similar pattern during its peak period to the Welwyn Up direction. Once again it can be seen that the Tariff Equivalents for the Fast lines are of a different character to those for the Slow lines. Whilst the Fast line Tariff Equivalents are again marked by clear peaks for the two infrastructure constraints (i.e. WEL and STH); the Slow lines have a single significant peak in the section between the two constraints (i.e. WST). This is again likely to reflect the presence of Langley Junction but in this case the Down Slow line has two-way traffic. It can be seen that following consolidation this peak is eliminated. However, the relatively high Tariff Equivalents for the two constraints (WEL and STH) under both Fast line and Slow line conditions obviously remain important following consolidation.

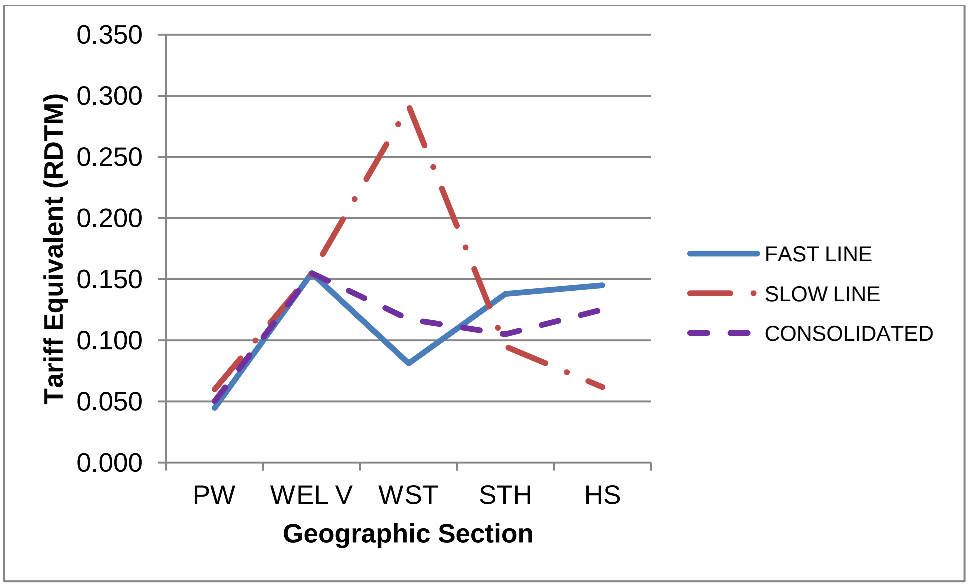


**Figure 8.3** Comparison of Fast and Slow Line Tariff Equivalents for the Welwyn Down Area During the Evening Peak Period (1600 to 1900).

To summarise the results shown in Figures 8.2 and 8.3, it can be seen that following consolidation the sections most associated with the infrastructure constraints attract high Tariff Equivalents. The single departure from this rule (i.e. the section in both directions between Stevenage and Woolmer Green Junction) contains a node within the section (i.e. Langley Junction) that has not been specifically modelled and also lies between the two infrastructure constraints that have. The consolidated patterns of Tariff Equivalents seen in Figures 8.2 and 8.3 therefore still reflect the interaction between traffic and infrastructure for both directions of traffic in the Welwyn area.

## 8.4 Discussion of Calculated Tariff Equivalents

As noted previously no adjustment has been made for the Infrastructure Fault Rate, the Schedule 8 Payment Rate or the Delay:Lateness Ratio. The Tariff Equivalents presented in this Chapter are therefore calculated in terms of Congestion Related Reactionary Delay per Train Mile (RDTM). The Tariff Equivalents also represent cost per individual train.

### 8.4.1 Tariff Equivalents for Individual Geographic Sections

This section illustrates calculated Tariff Equivalents for individual locations. These are presented for both three-hourly and hourly time bands.

Figure 8.4 shows the calculated sample Tariff Equivalent for the Welwyn Viaduct constraint in the Down direction for the six consolidated time periods.

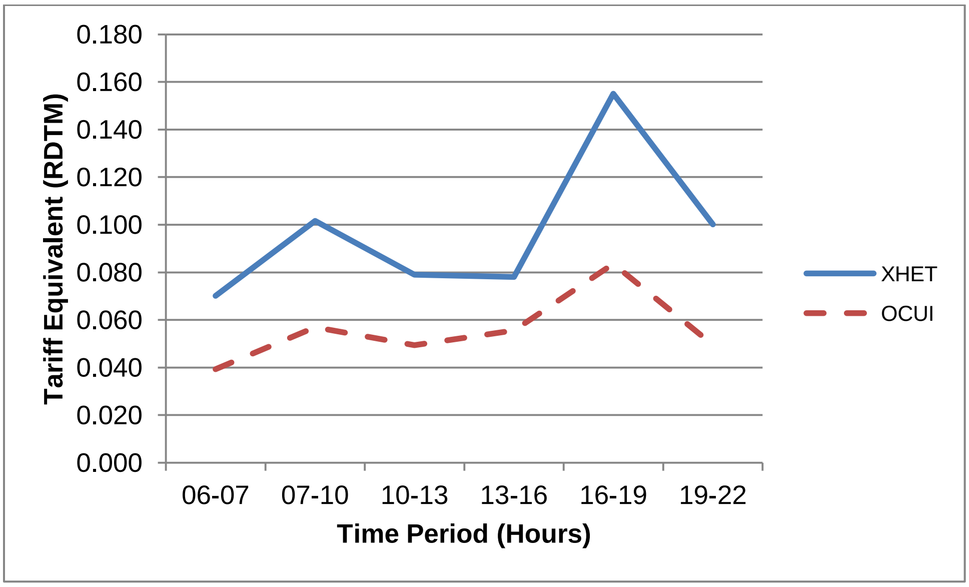


**Figure 8.4** Sample Tariff Equivalent for Welwyn Viaduct (Down) Using Averaged Three-hour Time Periods.

It can be seen that the Tariff Equivalents of the two capacity utilisation measures plotted show the same general pattern. There is a small ‘peak’ in the charge in the 0700 to 1000 hours time period and a more significant ‘peak’ in the charge in the 1600 to 1900 hours time period. These are consistent with the concept of charging more during the most congested part of the day on this section (i.e. the evening peak between 1600 and 1900 hours) but also distinguishing between capacity utilisation in other time periods (i.e. the morning contra-peak between 0700 and 1000 hours).

However, the graph also shows that the XHET Tariff Equivalents have a greater range than the OCUI ones. For example, the increase in the Tariff Equivalent for XHET between the 1300 to 1600 hours and 1600 to 1900 hours time periods is 0.077 compared to 0.028 for OCUI. The XHET Tariff Equivalent is also much higher than the OCUI one meaning that the cost of congestion is significantly greater with a HET based approach than a CUI based one.

In order to examine the impact of averaging on the calculated Tariff Equivalents, Tariff Equivalents based on hourly time periods were also calculated. Figure 8.5 shows the calculated hourly Tariff Equivalents for the Welwyn Viaduct (Down) section.

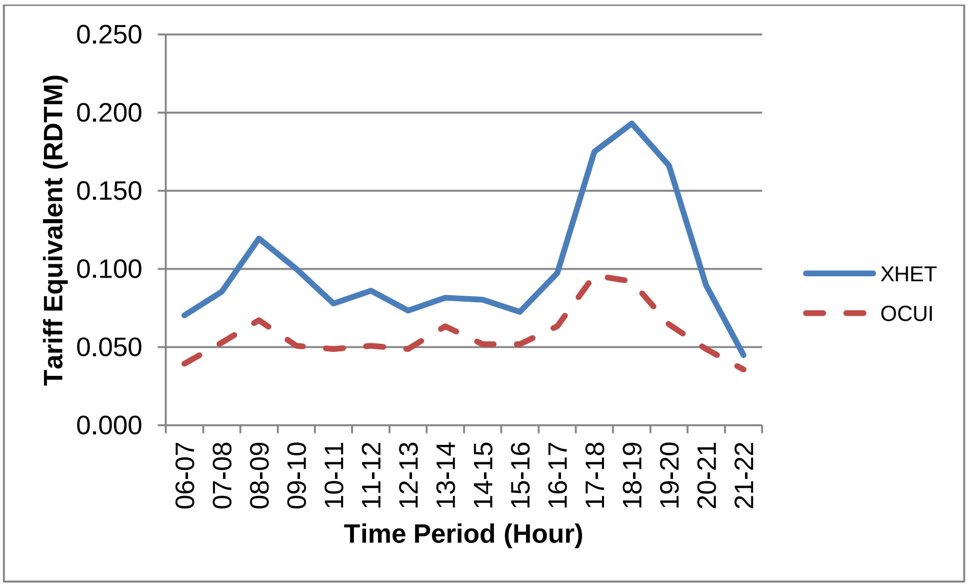


**Figure 8.5** Sample Hourly Tariff Equivalents for Welwyn Viaduct (Down).

It can be seen that the same peak and contra-peak pattern exists for the hourly Tariff Equivalents as the averaged three hourly Tariff Equivalents. Once again, the size of the difference between the various time periods, and hence the size of the incentive, is considerably greater with the XHET measure.

Figure 8.6 shows the three-hourly Tariff Equivalents for the Grantham to Newark geographic section. Once again XHET produces substantially higher Tariff Equivalents than OCUI. The XHET based Tariff Equivalents also show a very distinct and substantial evening peak compared with the OCUI value.

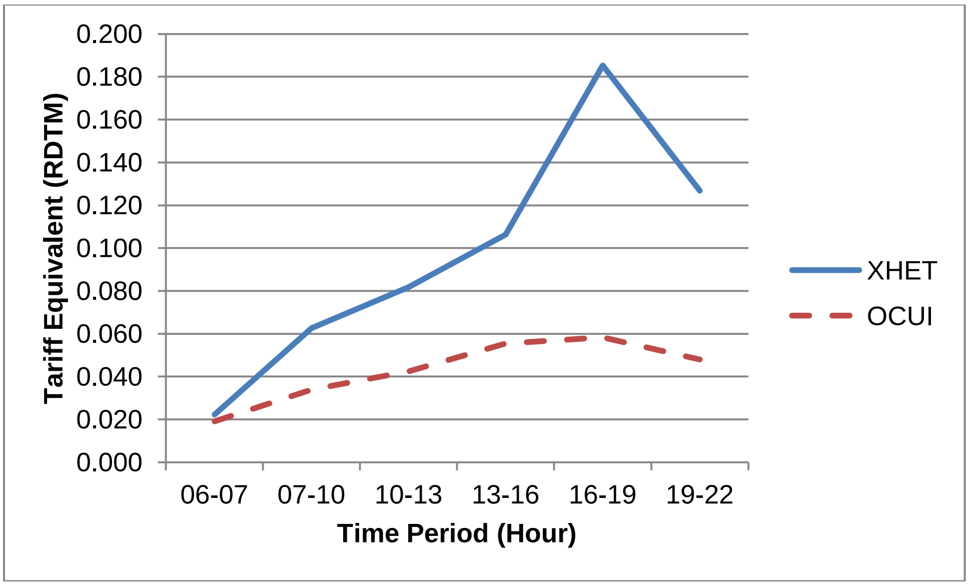


**Figure 8.6** Three-Hourly Tariff Equivalents for the Grantham to Newark (Down) Geographic Route Section.

Figure 8.7 shows the hourly Tariff Equivalents for this section. It can be seen that the variation between adjacent time periods is more extreme than those observed in Figure 8.4 (the hourly Tariff Equivalents for Welwyn Viaduct in the Down Direction). However, once again the XHET based Tariff Equivalents are substantially higher than the OCUI based ones and have more defined ‘peaks’ and ‘troughs’.

Although there is a definite evening peak (1700 to 1800) with XHET, there is not a significant morning peak and this combined with significant peaks at other times of the day produces an unusual profile. This reflects the fact that in the Newark Area capacity utilisation results from the mixture of passenger and freight traffic. This contrasts with the changes in the volume of passenger trains (with the occasional freight train) seen in the Welwyn area, which leads to more traditional peaks and inter-peak periods. Comparing the Tariff Equivalents in Figure 8.6 with those in Figure 8.7, shows that the utilisation of three-hour time periods leads to a definite smoothing of the charges produced.

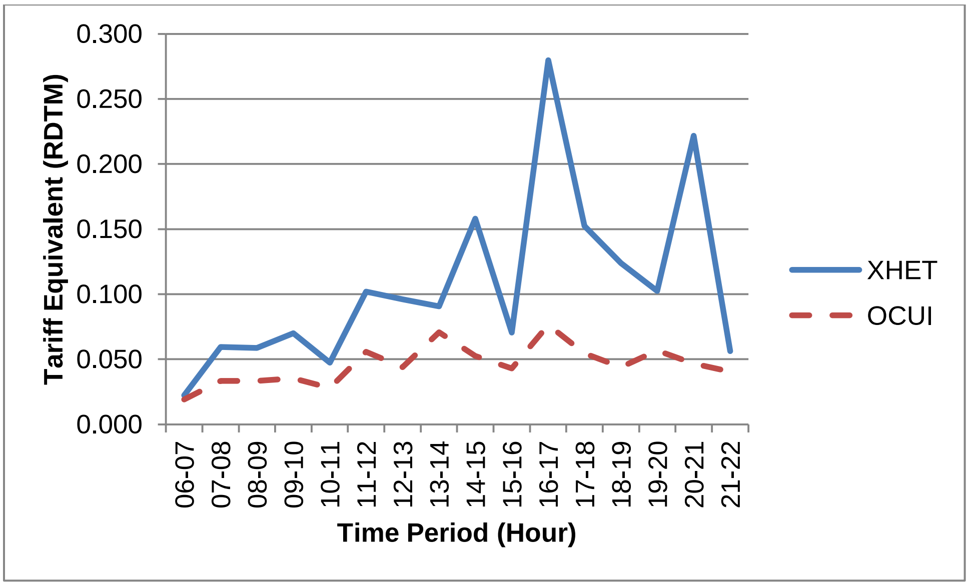


**Figure 8.7** Hourly Tariff Equivalents for the Grantham to Newark (Down) Route Section.

The observed profiles also demonstrate that the ‘slope’ of the relationship between capacity utilisation and reactionary delay is much ‘steeper’ for XHET than OCUI. This means (especially if combined with hourly Tariff Equivalents) that there is a greater incentive for traffic to be planned during less congested times if the HET based measures are used. Since the XHET Tariff Equivalents are also much higher than the OCUI ones and the utilisation measures themselves have been demonstrated to be more effective; a further conclusion can be drawn i.e. that the use of OCUI undercharges for congestion.

### 8.4.2 Combined Unweighted Area Tariff Equivalents

The next section compares the individual unweighted Tariff Equivalents for the  geographic sections within their respective areas.

Combining the sectional Tariff Equivalents into their respective areas shows that there are many similarities but some differences between the patterns produced by the two different capacity utilisation measures. Figure 8.8 plots the calculated Tariff Equivalents for the different geographic sections that form the Welwyn Up Area for the three-hour morning peak period (0700 to 1000 hours). The graph shows that the capacity utilisation measures produce Tariff Equivalents with the following characteristics:-

- A Medium Tariff Equivalent on the Sandy to Hitchin section (SH) i.e. in advance of the Hitchin Junction constraint for  XHET but a low Tariff Equivalent for OCUI. The difference may reflect the fact that the former measure accounts for junction capacity utilisation but the latter does not.

- Low Tariff Equivalents on the Hitchin to Stevenage section (HST) i.e. following the Hitchin Junction constraint.
- High Tariff Equivalents on the Stevenage to Woolmer Green Junction section (STW) i.e. in advance of the Welwyn Viaduct constraint.
- High Tariff Equivalents on the Welwyn Viaduct constraint itself (WEL).
- Low Tariff Equivalents on the Welwyn to Potters Bar section (WP) i.e. following the Welwyn Viaduct constraint.

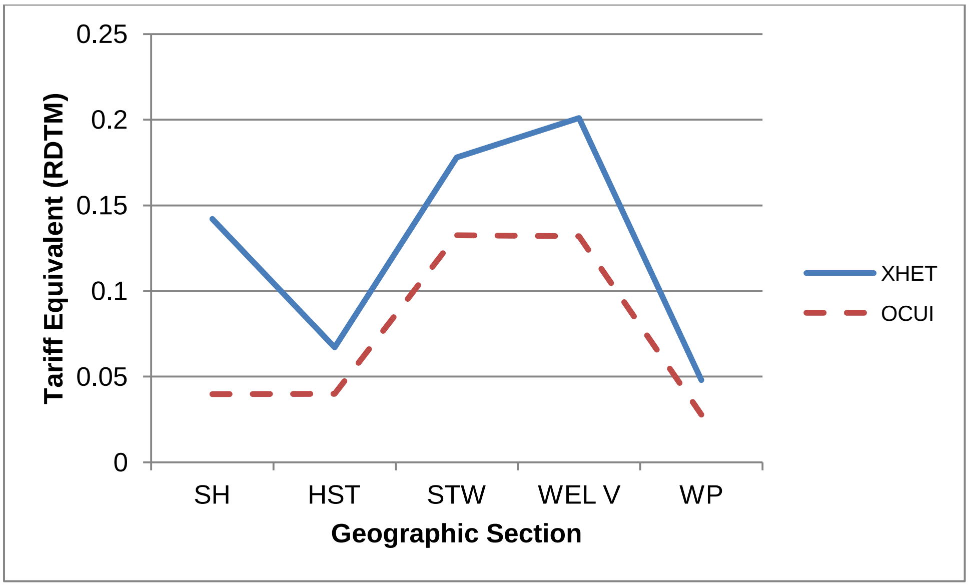


**Figure 8.8** Comparison of the Calculated Tariff Equivalents for the Geographic Sections that form the Welwyn ‘Up’ Area (0700 to 1000 Hours).

Figure 8.8 shows that the level of Tariff Equivalents in the Welwyn ‘Up’ Area is related to the presence of infrastructure constraints. This is in line with expectations that the incentive effect of tariffs should reflect the greatest potential congestion. This is particularly true for the Welwyn Viaduct constraint but applies in the case of XHET to the Hitchin Junction constraint. It can be seen that the sections in advance of constraints generally attract greater Tariff Equivalents than those following them. In the case of Welwyn Viaduct the difference is considerable. Although, it could be argued that this arises due to the general policy of including the capacity utilisation of junctions in the link immediately in advance of them; the pattern also applies to the link-only capacity utilisation measure (i.e. OCUI).

Figure 8.9 however presents a different picture in the case of XHET. It shows the calculated Tariff Equivalents for the three geographic sections that form the Newark ‘Down’ area for the 1600 to 1900 hours time period. It can be seen that the XHET Tariff Equivalent is highest by a substantial margin for the Newark to Grantham section (NG) , rather than the Retford to Newark section (RN). In contrast, the Tariff Equivalent for OCUI shows almost identical values for the two sections. This pattern seen with the XHET measure reinforces the idea that congestion (and therefore the charge for it) due to a mix of traffic can be as important as that due to the presence of physical infrastructure constraints.

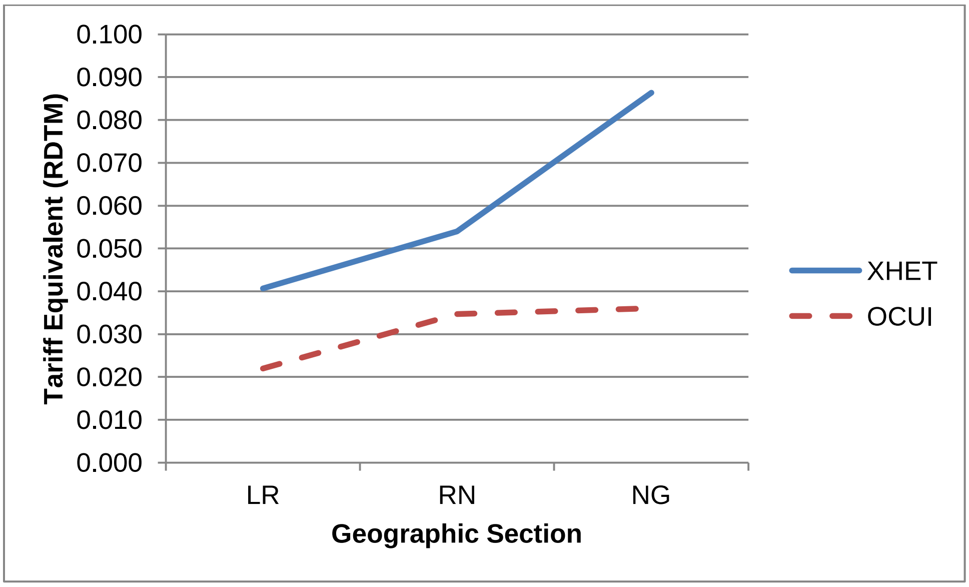


**Figure 8.9** Comparison of the Calculated Tariff Equivalents for the Geographic Sections that form the Newark Down Area (1600 to 1900 Hours).

### 8.4.3 Weighted Area Tariff Equivalents

The next section examines the effect of combining individual sections into their respective areas using the weighting methodology described earlier in this Chapter.

Whilst the previous graphs presented individual unweighted Tariff Equivalents for the sections that formed particular areas, Figure 8.10 shows the weighted (by section) Tariff Equivalents for the Welwyn ‘Up’ Area using XHET and OCUI. It can be seen that once again XHET produces higher

Tariff Equivalents than OCUI. The graph shows that XHET exhibits  as significant peaks in Tariff Equivalents for the 0700 to 1000 time period and 1900 to 2200 hours period. Although increases in the OCUI Tariff Equivalent can also be seen for OCUI for the same periods, these are much less substantial. Overall OCUI exhibit much less variation during the day in the size of the Tariff Equivalents than can be seen with the XHET ones. Once again it can be concluded that this pattern arises due to the greater regression 'slope' seen with XHET.

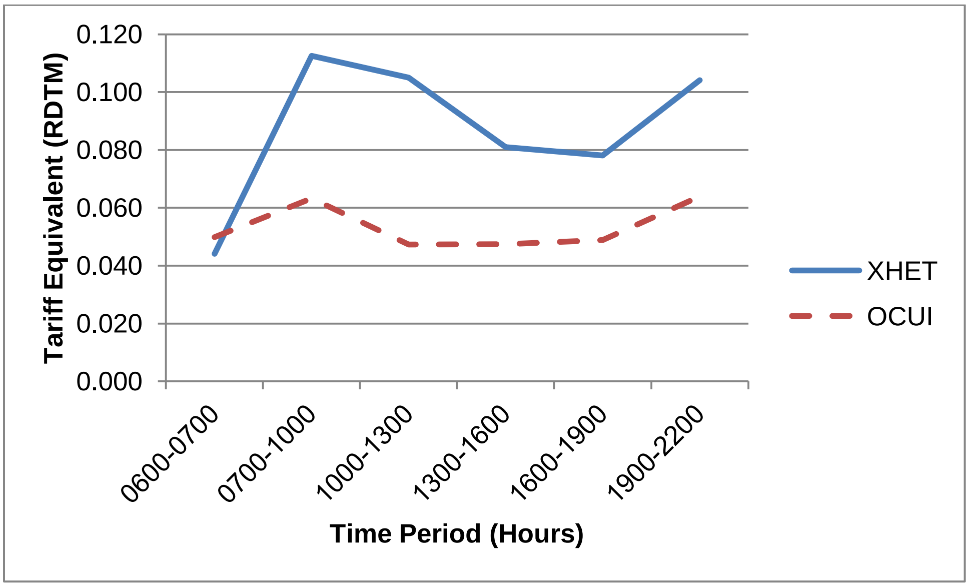


**Figure 8.10** Overall Tariff Equivalent Per Train Mile for the Welwyn 'Up' Area (Using the Weighted Sectional Results).

Figure 8.11 shows the weighted sectional results by time period but in this case for the Newark Down area.

It can be seen that for both capacity utilisation measures there is a peak in the size of the Tariff Equivalent between 0700 and 1000 hours and then a much bigger peak between 1900 and 2200 hours. Once again, the XHET measure produces the highest Tariff Equivalent with more distinct variation in the profiles of the charges than seen with OCUI.

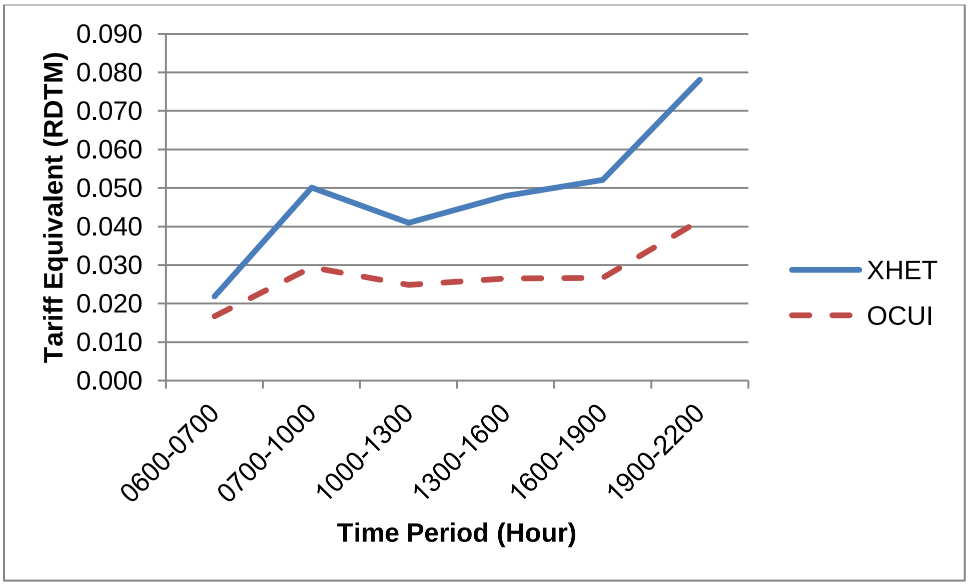


**Figure 8.11** Overall Tariff Equivalent Per Train Mile for the Newark ‘Down’ Area (Using the Weighted Sectional Results).

Both Figures 8.10 and 8.11 show that there are distinct time of day variations in the size of Tariff Equivalents. However, as noted these differences are markedly greater with the XHET capacity utilisation measure.8.4.4 Service Code Tariff Equivalents

The final element that needs to be considered is the consolidation of Tariff Equivalents by Service Code. This final step will then replicate the approach taken for the 2013 recalibration of the Capacity Charge as far as is possible. Once again for the sake of brevity results are only presented here for XHET (the ‘best’ performing sectional capacity utilisation measure) and OCUI (the sectional capacity utilisation measure closest to the one used in the ‘original’ analysis).

One useful aspect with the two geographic areas chosen is that although many of the freight and passenger trains are restricted to one or the other, long-distance passenger trains pass through both. This means that the Service Code Tariff Equivalents in their case will reflect elements of the marginal costs for both geographic areas.

Another helpful aspect is that the two areas can be coordinated quite easily with each other. For the regression analysis there was no direct connection between the Newark and Welwyn areas. However, it is possible to carry out a simple comparison of results for the two areas since the standard journey time between the two principle constraints (Newark Flat Crossing and Welwyn Viaduct) is almost exactly one hour. For example, the 1A13 Skipton to London Kings Cross path (which has only 0.5 minutes of pathing time between Loversall Carr and Potters Bar) has a planned time of 0832.5 at Newark Flat Crossing and 0930.5 at Woolmer Green Junction. This means that one hour for the Newark Area effectively equates to the next hour for the Welwyn Area in the Up direction. In other words, the 0600 to 0700 hours period in the Newark Up area can be compared with the 0700 to 0800 hours period in the Welwyn Up Area. This connection allows sample Tariff Equivalents based on Service Codes to be calculated.

**Table 8.2** Sections Used to Create Tariff Equivalents by Partial Service Code (SC) for the Up Direction.

| Operator | Areas | Geographical Sections | SC Equivalent |
|---|---|---|---|
| East Coast Trains | Newark | Loversall Carr to Grantham<br>Sandy to Potters Bar | 1<br>(LD) |
| Hull Trains | Newark<br>Welwyn | Loversall Carr to Grantham<br>Sandy to Potters Bar | 1<br>(LD) |
| Grand Central | Newark<br>Welwyn | Loversall Carr to Grantham<br>Sandy to Potters Bar | 1<br>(LD) |
| Freight | Newark | Loversall Carr to Grantham | 2<br>(FT) |
| GN (Peterborough) | Welwyn | Sandy to Potters Bar | 3<br>(PBO) |
| GN (Cambridge) | Welwyn | Hitchin to Potters Bar | 4<br>(CAMB) |

The Service Code equivalents[44] used in the analysis are shown in Table 8.2. For the purposes of this analysis these have been simplified. So for example, all East Coast trains, Grand Central and Hull Trains services are combined into a single ‘long-distance’ (LD) Service Code. This reflects the

[44] Although complete Service Code data has not been analysed , combining the two different areas to produce partial Service Codes gives an idea of the process and outcome. This allows further discussion of potential tariff mechanisms.

fact that these three operators use the same geographic sections and areas (i.e. the entirety of the Newark and Welwyn areas). The freight  (FT) Service Code Tariff Equivalent is just the Newark area reflecting the limited volume of freight in the Welwyn area. It can be seen that in total four different Service Code Tariff Equivalents are created. The sections which form each Tariff Equivalent are given in the table.

As illustrated in Example Five (Figure 8.1) these Service Code Tariff Equivalents were calculated using a weighted mileage approach.

Figure 8.12 shows hourly[45] Service Code Tariff Equivalents for the four groupings for the XHET results. It can be seen that each of the Service Code Tariff Equivalents for the most part lie within the range of 0.050 and 0.100 RDTM.  However, the GN (Peterborough) Service Code Tariff Equivalent produces the highest level for almost the entire day whilst the GN (Cambridge) Tariff Equivalent at times produces the lowest. This is perhaps unexpected given that both types of services pass through the Hitchin Junction and Welwyn Viaduct constraints.  Analysis of the base data set shows that the explanation for this lies with the Sandy to Hitchin geographic section. The Fast and Slow lines generate relatively high levels of reactionary delay for relatively low levels of traffic. This translates into an increased Tariff Equivalent for the Peterborough flow compared to the Cambridge flow. This example underlines a key issue with the use of Service Code based tariffs i.e. they are sensitive to each section of the ‘journey’ that comprises the flow. Of course , whilst the Sandy to Hitchin link is present in the data set there is no equivalent link on the Cambridge branch adjacent to the junction.

Each of the Service Codes has noticeable morning and evening peaks. However, there are some important differences between the profiles of the four tariffs.  For example, the morning peak is at different stages during the day for the different types of passenger traffic. It can be seen that the two GN (commuter) Service Codes have a peak at 0800 to 0900 hours. However, the Long Distance service code peak is much later at 1000 to 1100 hours. In combination the two individual peaks provides an explanation for the extended morning peak seen in the sample timetable.

[45] The methodology used to combine the Newark and Welwyn areas means that only the period 0700 to 2200 hours can be considered in this part of the analysis.

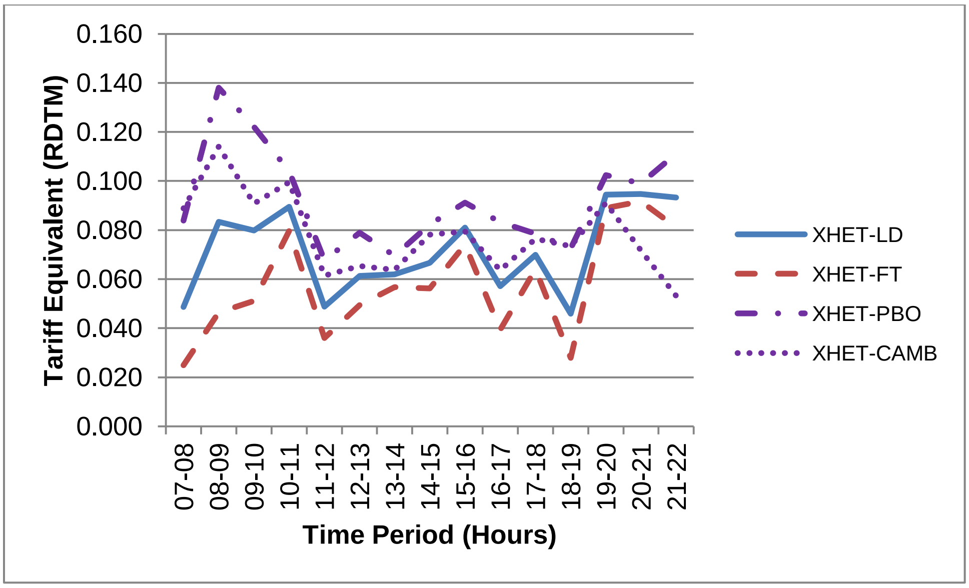


**Figure 8.12** Service Code Hourly Up Tariff Equivalents (XHET).

There are several reasons for the later long-distance peak. Firstly, as described in Chapter Six, East Coast services include those with origins as far away as Scotland. Average travel time between Edinburgh and London is approximately 4.5 hours. A 0600 Departure from Edinburgh would therefore arrive in London at approximately 1030 hours. This later morning peak therefore reflects the arrival of some of East Coast’s first trains in the sample network. Secondly, the services of the two open access operators have their initial services planned outside the ‘recognised’ peak hours. The later long-distance peak reflects the paths operated by these companies. Finally, it can be seen that the morning peak for the freight Tariff Equivalent is also at 1000 to 1100 this therefore reflects the nature of congestion in the Newark area.

The pattern of the four Tariff Equivalents also differs in other parts of the day. It can be seen that the GN Service Code Tariff Equivalents have reasonably similar profiles although as discussed earlier the scale of them are very different. However, the Long-distance passenger and Freight Tariff Equivalents look even more similar . It is clear that the inclusion of the Newark area Tariff Equivalents which are more reflective of the mix of traffic in the timetable rather than the volume of traffic has a profound effect on the calculated Service Code Tariff Equivalents. This conclusion is reinforced by the existence of additional peaks during the day for the long-distance passenger and the freight Service Codes.

Figure 8.13 consolidates the XHET combined Tariff Equivalents into three-hourly ones. It can be seen that although much of the definition obtained from an hourly Tariff Equivalent is lost, there are still discernible patterns. Once again the Peterborough Service Code has the highest Tariff Equivalent associated with it. Both the Peterborough and the Cambridge Tariff Equivalent have clear morning peaks. However the Long-distance Tariff Equivalent and freight Tariff Equivalent do not.  Instead there is a discernible peak in the 1300 to 1600 time period for the Freight Service Code and a significantly larger peak for both Service Codes in the 1900 to 2200 hours period.

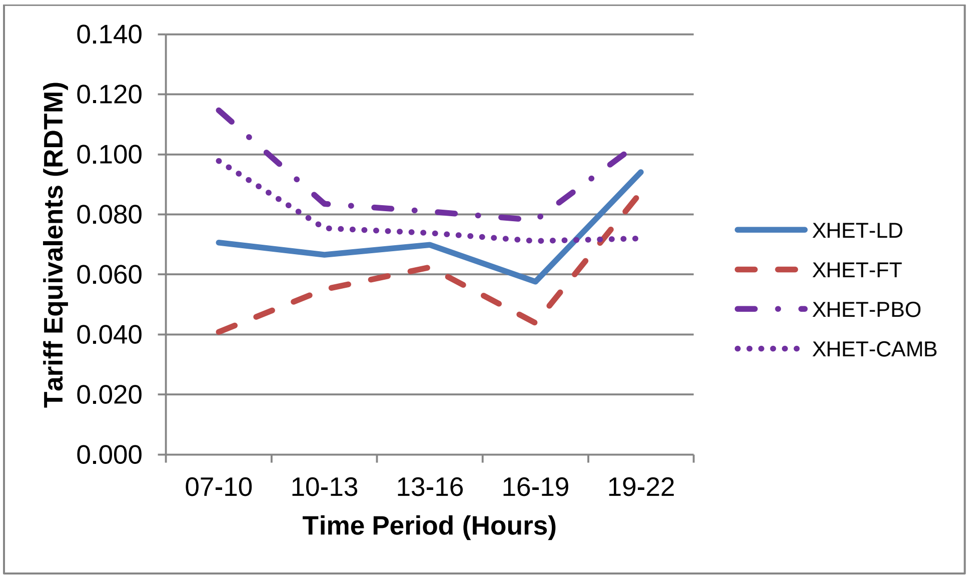


**Figure 8.13** Service Code Three Hourly Up Tariff Equivalents (XHET).

This ‘spike’ for the Freight Service Code during the ‘inter-peak’ is contrary to expectations and once again leads to the conclusion that the Tariff Equivalents  are impacted by the inclusion of the Newark area. The implications of this on charging for congested infrastructure will be discussed later in the chapter.

Figure 8.14 presents the hourly Tariff Equivalents for the four Service Codes using the OCUI measure.

It can be seen that the overall Tariff Equivalents are less than those presented in Figure 8.12. The OCUI Tariff Equivalents follow the standard pattern of being considerably lower than the XHET Tariff Equivalents and in-fact they are approximately half the level. The distinguishing features of the XHET diagram can also be seen here, for example the later morning peak for the long-distance and freight service codes, are still visible. However, the lower values means that the variation observed in Figure 8.12 is much less.

This reduced variation means that there is less financial incentive to plan trains at less congested times of the day.

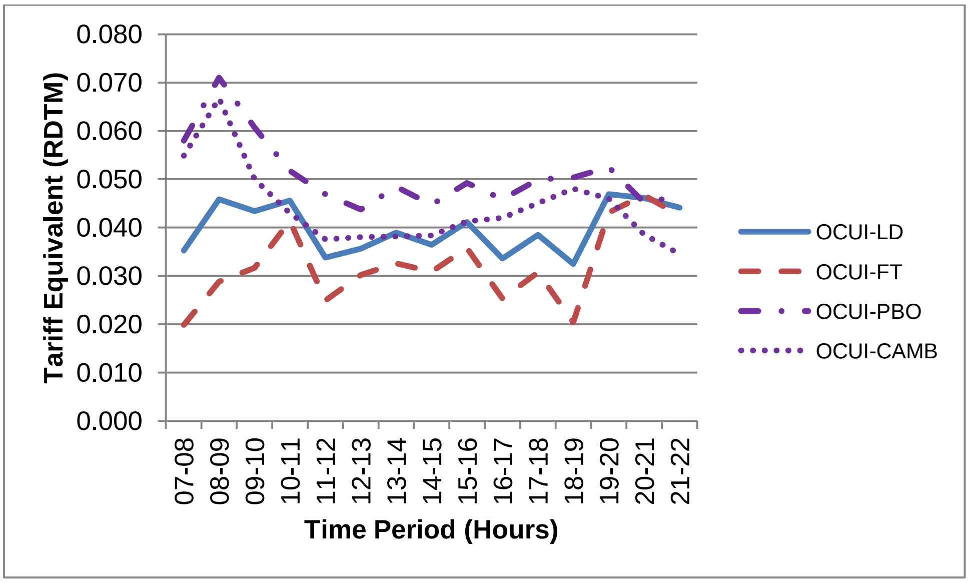


**Figure 8.14** Service Code Hourly Up Tariff Equivalents (OCUI).

One difference is that the Cambridge Service Code Tariff Equivalent has a high morning peak value comparable with that for the Peterborough Service Code. This was not seen with the XHET results. Analysis of the data shows that OCUI calculates a much lower utilisation on the Sandy to Hitchin Up Slow between 0700 and 1000. This is because OCUI does not take into account the impact of the Hitchin junction moves (as well as any timetable bunching). This means that the Peterborough Service Code Tariff Equivalent is lower than seen with XHET.

Figure 8.15 shows the consolidated three-hourly 'Up' Tariff Equivalents using the OCUI results. Once again the OCUI Tariff Equivalents are approximately half their XHET counterparts. It can be seen that there is again much less variation. However, this does mean that some of the less logical aspects of the pattern seen in Figure 8.13, e.g. the inter-peak 'spike' for the freight service codes is not apparent.

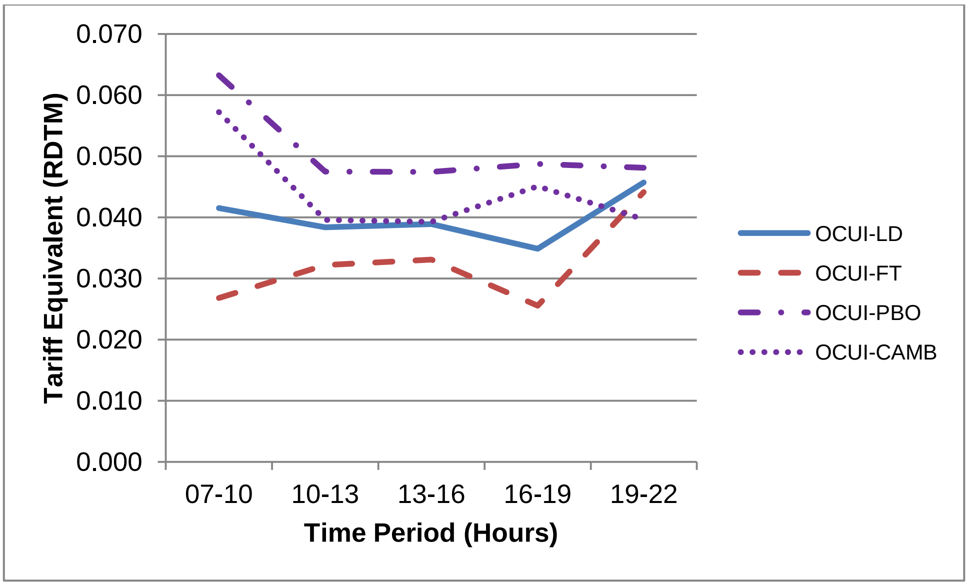


**Figure 8.15** Service Code Three Hourly ‘Up’ Tariff Equivalents (OCUI).

Table 8.3 directly compares the three hourly up Tariff Equivalents using the XHET and OCUI capacity utilisation measures. For the sake of brevity only the Long-distance and GN Peterborough service codes are given.

**Table 8.3** Comparison of the Three Hourly consolidated Tariff Equivalents for the ‘Up’ direction.

| Time Period | Long-Distance XHET (RDTM) | Long-Distance OCUI (RDTM) | Peterborough XHET (RDTM) | Peterborough OCUI (RDTM) |
|---|---|---|---|---|
| 0700-1000 | 0.071 | 0.041 | 0.115 | 0.063 |
| 1000-1300 | 0.067 | 0.038 | 0.084 | 0.047 |
| 1300-1600 | 0.070 | 0.039 | 0.081 | 0.047 |
| 1600-1900 | 0.058 | 0.035 | 0.078 | 0.049 |
| 1900-2200 | 0.094 | 0.046 | 0.104 | 0.048 |

The much greater size of the XHET Tariff Equivalents can be clearly seen. The table also illustrates the difference between the ‘Peterborough’ Tariff Equivalents and the ‘Long Distance’ passenger Tariff Equivalents. As noted earlier, the Peterborough Tariff Equivalents are higher than the Long-distance Tariff Equivalents and have a clear morning peak of 0700 to 1000 hours. However, one aspect to bear in mind when considering the difference in size between the two Service Codes is that the long-distance trains have much longer journeys. The total Tariff Equivalent levied on each long-distance train will therefore be greater than that levied on each Peterborough train.

Table 8.4 shows each of the Tariff Equivalents consolidated into a single value that covers the entire modelled day. This was simply calculated by averaging the obtained hourly tariffs for each of the Service Codes. The greater cost of the XHET tariffs can be clearly seen. The differing values between the various Service Codes are also clear.

**Table 8.4** Comparison of the Daily (0700 to 2200 hours) Consolidated Tariff Equivalents for the 'Up' Direction.

| Service Code | XHET Tariff Equivalent (RDTM) | OCUI Tariff Equivalent (RDTM) |
|---|---|---|
| Long-distance passenger (LD) | 0.0718 | 0.0399 |
| Freight (FT) | 0.0579 | 0.0323 |
| GN (Peterborough) | 0.0923 | 0.0510 |
| GN (Cambridge) | 0.0781 | 0.0442 |

### 8.4.5 Summary

A number of important observations can be made about the Tariff Equivalents that were calculated as part of the analysis for this thesis:-

- The more 'affective' HET based capacity utilisation measures produce higher Equivalent Tariffs than the CUI based capacity utilisation measures. This reflects the steeper 'slope' (using the β value)[46] associated with the relationship between HET capacity utilisation and reactionary delay. The implication is that the CUI based Tariff Equivalents under-charge for congestion.

[46] As outlined in Section 5.4 the tariffs reflect the difference between two points on the regression line which represents the capacity utilisation of an 'additional' train. The steeper line associated with HET produces higher tariffs.

- The more ‘affective’ HET based capacity utilisation Tariff Equivalents produce a greater variation in value than the CUI based capacity utilisation measures. In other words there is greater contrast and thus incentive between time periods. Once again this is a reflection of the steeper ‘slope’ associated with the HET based measures relationship with reactionary delay. The implication is that the CUI based Tariff Equivalents do not differentiate sufficiently between the costs of congestion at different levels.
- The impact of three-hour Tariff Equivalents compared to one-hour Tariff Equivalents produces as expected a ‘smoothing’ effect. In other words the incentive impact is lessened by the use of consolidated Tariff Equivalents.
- Examining Tariff Equivalents for individual sections within each area shows that the presence of infrastructure constraints is reflected in the calculated tariffs for the Welwyn area but not the Newark Area. The pattern of tariffs for the latter is instead influenced by the timetable constraint in the area (i.e. the significant heterogeneity of traffic).
- One aspect not previously discussed is that the Tariff Equivalents derived from the sectional capacity measures are higher than those derived from the area capacity measures. The calculated marginal cost of congestion is therefore higher. Given that the sectional Tariff Equivalents are considered more affective, it can be concluded that the area Tariff Equivalents under-charge for the cost of congestion.
- The Tariff Equivalents for different Service Codes alter according to whether the paths pass through the Newark area as well as the Welwyn one. The long-distance Tariff Equivalents which pass through both have a significantly different character to the those that are just based on the Welwyn area. This difference could lead to unwanted patterns of behaviour where operators are in-fact penalised less for new services at more congested parts of the day. For example, Table 8.3 shows that East Coast, Hull Trains and Grand Central would pay their lowest XHET Tariff Equivalent for services during the traditional evening peak (i.e. 1600 to 1900 hours) and one only marginally higher than the off inter-peak period for the traditional morning peak (i.e. 0700 to 1000 hours). Although, there is a reason for this (i.e. the later peak for long-distance traffic) it is illogical that these services should actually be incentivised to operate in the Welwyn area during its most congested periods.

Adopting the approach used to produce Britain's Capacity Charge to create suitable congestion tariffs, can therefore lead to reduced incentives to operate in less congested times. In some cases the Tariff Equivalents appear to be illogical (as noted in the final point above). The adoption of a Service Code approach also removes the direct link to the location of the congestion, as each train service is charged at a flat rate per mile.

## 8.5 Alternative Thoughts on the Introduction of Tariffs

### 8.5.1 Overview

The theoretical evidence supported by the findings of this thesis suggests that reactionary delay increases at an exponential rate as congestion rises. The idea that Network Rail should receive some form of compensation for the increased marginal cost of congestion does seem appropriate if they are to be encouraged to optimise the volume of traffic that makes use of the British Rail network.

The results of the analysis described in Chapter Seven suggests that the most appropriate course of action would be to calculate a tariff based on the XHET capacity utilisation measure. However, there is a potential problem with this. As noted earlier, a key difference between CUI and HET is that the former assumes traffic with identical characteristics are evenly spaced whilst the latter accounts for their actual planned spacing in the timetable. The HET based Tariff Equivalents are higher than the CUI based ones. Network Rail therefore would receive more compensation for the marginal cost of congestion. This compensation can theoretically be maximised by Network Rail revisiting the timetable and increasing the level of even spacing thus reducing the amount of reactionary delay[47]. The 'poorer' the base timetable the greater the potential level of compensation. In other words, Network Rail is effectively being rewarded for carrying out one of their duties which is the production of efficient timetables. The acceptability of this possibility is discussed in this section.

Furthermore, the previous sections in this chapter have shown that greater detail produces more precise Tariff Equivalents and thus theoretically clearer price signals. The 'smoothing' impact of adopting three-hourly time periods

[47] This comment is based on the fact that timetables can change substantially every six months whilst the intention is that the Capacity Charge would only change each Control Period (or 5 years) to maintain transparency and stability of the charges.

and the further effect of the adoption of a Service Code approach has been discussed. It is worth remembering though that the Service Code tariff system was adopted due to the difficulties and costs associated with billing anything more disaggregated (Arup, 2013).

A final issue which was raised previously in this thesis is the fact that the tariff applies to all trains in all periods whether or not a new service has been introduced. Although, the ORR agreed that this was the most appropriate way to levy the Capacity Charge; it is still worth considering this aspect again during the discussion within this Chapter.

This section therefore considers the following aspects that need to be considered as part of any recommendation for an alternative charging mechanism:-

- How to apply the HET based measures to ensure that Network Rail is compensated for the increased marginal cost of introducing new services on the network rather than rewarded for behaviour that is already expected (i.e. the optimisation of timetables and the efficient use of capacity).
- How tariffs should geographically be applied. Previously in this Chapter the implication has been that a tariff will apply to all sections. However, as described earlier in this thesis the Theory of Constraints suggests that the efficiency of a system is dictated by the efficiency of its principle constraints. Although, it has been demonstrated that capacity utilisation measures based on the theory are not particularly effective there is a case that the tariffs themselves should only apply to the most congested parts of the network.
- To what time periods, if any, should congestion charges be applied to? As discussed earlier in this Chapter, the increase in time periods from hourly to three hourly smoothes the Tariff Equivalent and thus reduces the effectiveness of the incentive.
- Linked to the second and third points is the question of whether there could be a greater disaggregation of tariffs beyond that of the Service code that would potentially avoid the issues associated with billing and implementation raised by Network Rail (Arup, 2013).
- Should tariffs just apply to new additional traffic or should tariffs as now apply to all traffic?

- Should there be a link between the Capacity Charge and the Declaration of Congested Infrastructure?

### 8.5.2 An Alternative Application of the Capacity Utilisation Measures

An option would be to use even spacing based on the HET methodology to calculate tariffs. In this case the percentage to be applied would follow the assumption that all the trains in the time period (and not just the additional one) were perfectly evenly spaced. The results of this approach are shown in Figure 8.16.

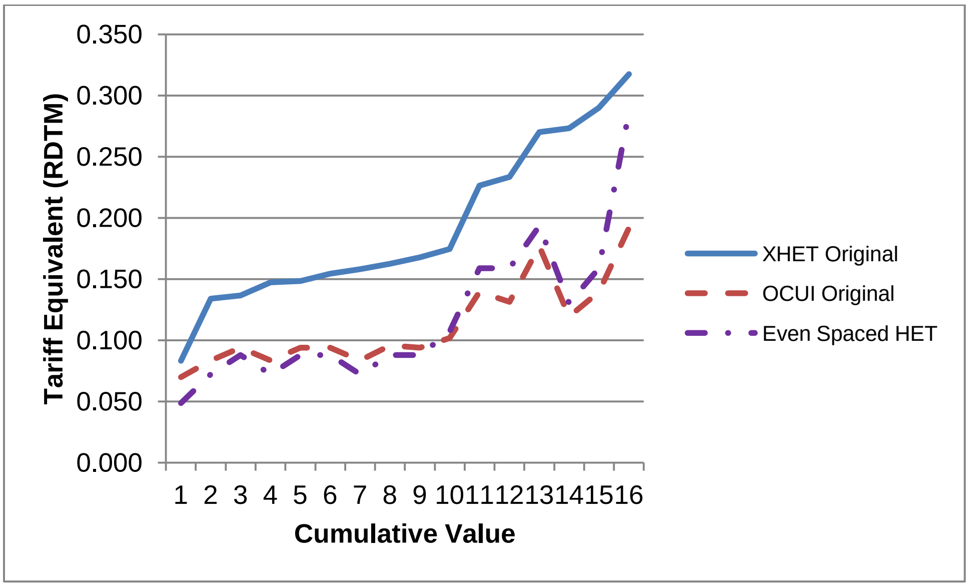


**Figure 8.16** Comparison of ‘Original’ and ‘Evenly Spaced’ Tariff Equivalents for the Grantham to Newark Geographic Section.

Figure 8.16 shows two original Tariff Equivalents and the possible new evenly-spaced XHET Tariff Equivalent for the Grantham to Newark geographic section. This section was chosen due to the ‘mixed’ nature of the traffic and therefore it provides a good example of the key issue of applying even-spacing to an irregular spaced timetable. The data is presented cumulatively so that the increase in the Tariff Equivalents can be compared.

It is worth noting that the original XHET line is considerably steeper than the CUI line. Once again it can be seen that HET suggests a much greater exponential relationship between capacity utilisation and reactionary delay than CUI. On Figure 8.16 the Evenly Spaced HET Tariff Equivalent line is generally similar to the OCUI line, until the highest level of capacity utilisation is reached. At this point there is a dramatic rise in the Tariff Equivalent which produces a level closer to the original XHET line.

The adoption of an evenly spaced HET approach therefore appears to provide the necessary compromise between adequate levels of compensation for Network Rail at high levels of congestion but not too significant levels of reward when capacity utilisation is more stable.

### 8.5.3 Geographical Application of HET Based Tariffs.

Superficially the application of tariffs to just the most congested parts of the rail network is an attractive option. Operators would be encouraged to plan services via less congested routes or contribute to the marginal cost if they continued to operate through the 'bottlenecks'. This approach matches the theory and application in several cases of road pricing. However, this ignores one of the fundamental differences between private car use and publically accessible rail travel. Whilst the driver of a car is free to choose the most appropriate route between their origin and destination, taking into account cost and journey times; franchised passenger train operators are not. Franchises dictate the route and intermediate stops that passenger operators must adhere to. Routes serving large 'markets' will naturally be busy. Although, Open Access operators have greater freedom it is logical that they will also wish to serve busy markets. This is the case for the two Open Access operators on the ECML. For freight traffic the situation is slightly different. In their case the choice of routing depends on the specific nature of the load and the origin and destination. Suitable alternative routes may simply not be available.

Any waiving of tariffs on less congested routes would also have to be very carefully managed. There are clearly risks with an ill thought out approach. As noted, franchised passenger operators would probably not be able to switch routes whilst open access operators would probably not wish to. There would however be an incentive for Freight operators to divert if the alternative route was suitable for the freight traffic in question and the other associated costs (such as any increased journey time) did not outweigh the savings in congestion charge costs. However, too great a transfer of traffic could firstly lead to the alternate route itself becoming congested either due to the increased volume or increased heterogeneity and secondly  could lead to more traffic on the primary route as passenger traffic (more able to pay the congestion charges) is attracted to fill the gaps left by the transferring freight traffic.  The objective is clearly that  the tariffs should encourage an equilibrium of traffic flows.

This thesis has also demonstrated that tariffs should differ within the actual sample areas of the ECML used for the analysis. For example, the impact of physical infrastructure constraints on reactionary delay and hence the level of Tariff Equivalents  has been shown to change between the Newark and Welwyn areas. Whilst, the presence of Hitchin Junction and Welwyn Viaduct is linked to increased reactionary delay on adjacent links in the Welwyn Area; Newark Flat Crossing has less of an impact than the mix of traffic in the timetable on reactionary delay in the Newark Area.

This is illustrated in Figure 8.17. The graph shows the calculated Tariff Equivalents for the Newark Up and Welwyn Up areas for the 0800 to 0900 adjusted period (i.e. the Newark sections are 0700 to 0800 hours and the Welwyn sections are 0800 to 0900 hours). The most striking aspect is the difference between the Newark area Tariff Equivalents and the Welwyn area Tariff Equivalents.

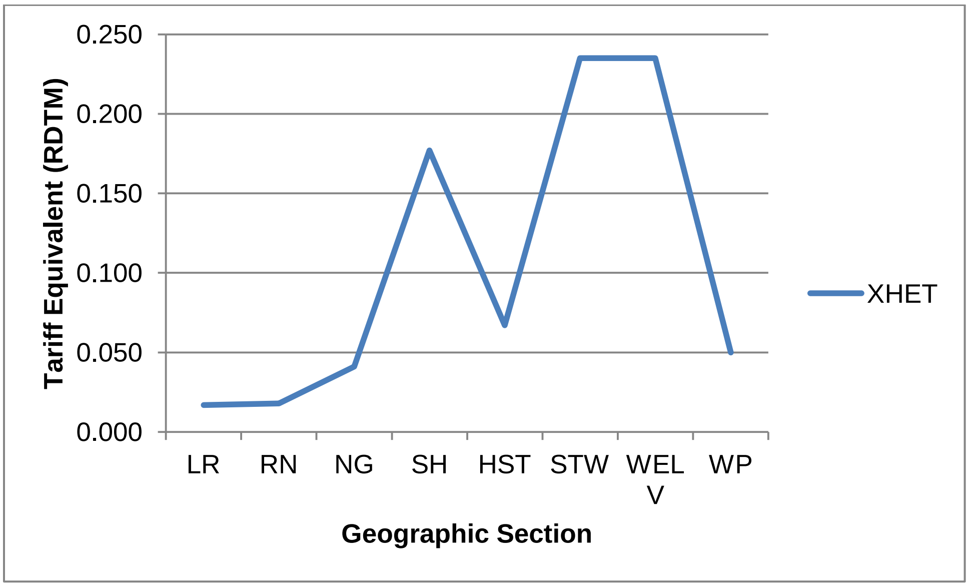


**Figure 8.17** XHET Tariff Equivalents for Each of the Up Geographic Sections (0800 to 0900 Hours).

Once again the Tariff Equivalents for the three sections that form the Newark Area do not reflect the presence of the physical constraint, i.e. Newark Flat crossing, which lies within the Retford to Newark section (RN). Instead the Newark to Grantham section (GN) has the highest Equivalent Tariff of the three sections. These Tariff Equivalents again demonstrate the importance of timetable heterogeneity in the Newark area. Since the congestion is more traffic driven than infrastructure driven, it can be argued that charging for only a limited number of sections is not feasible. In other words there is not a sufficient link in the Newark area between the nature of a geographic section and its level of congestion for consistent decisions to be possible about where to levy charges. In other words the results suggest that a Theory of Constraints approach to congestion charging is not to be recommended.

In contrast, as seen previously in this chapter, Tariff Equivalents in the Welwyn area reflect the presence of the two infrastructure constraints Hitchin Junction and Welwyn Viaduct. The approach to Hitchin Junction (i.e. SH) and the Viaduct itself (i.e. WEL V) have two of the highest Tariff Equivalents in Figure 8.17. This suggests that in this case it would be possible to levy a congestion charge for just these constraints. However, it can also been seen that the Tariff Equivalents for the section approaching Welwyn Viaduct (i.e. STW) is also very high. In addition, although low in comparison with the other Welwyn Tariff Equivalents the two remaining sections (namely HST and WP) which follow the constraints are still much higher than the section which contains Newark Flat Crossing (i.e. RN). In other words, levying charges solely based on the presence of constraints is not necessarily the most effective approach due to the importance of adjacent links. Solely charging the tariffs for the two actual infrastructure constraints in the Welwyn area would ignore their full impact on reactionary delay.

The reasons given above suggest that a Capacity Charge (or congestion charge) should continue to be levied for each geographic cell in the rail network. However, there is still the question of whether the charge should continue to be consolidated into a single charge per Service Code. The use of Service Code charges, or in the case of freight traffic one single charge, although being easy to administer does reduce the level of incentive. Freight traffic in particular which may have a greater ability to be rerouted receives no incentive to operate on less congested routes. Figure 8.18 shows the reduced level of incentive produced by consolidating the Tariff Equivalents shown in Figure 8.17 into a single charge per mile.

The graph clearly shows that the significant peaks and troughs obtained from a charging regime disaggregated to a geographic cell level disappear at the Service Code level. This is acceptable for traffic providing it passes through each of the geographic sections shown since the charge is calculated using the weighted average of each link. However, problems will arise for passenger Service Codes that contain trains operating on different routes. It is certainly a problem for freight traffic with its single charge despite the enormous variety in the nature of the traffic on the British rail network.

As discussed previously though, having individual charges for Freight traffic for each of the geographic sections that form the British rail network seems totally impracticable.

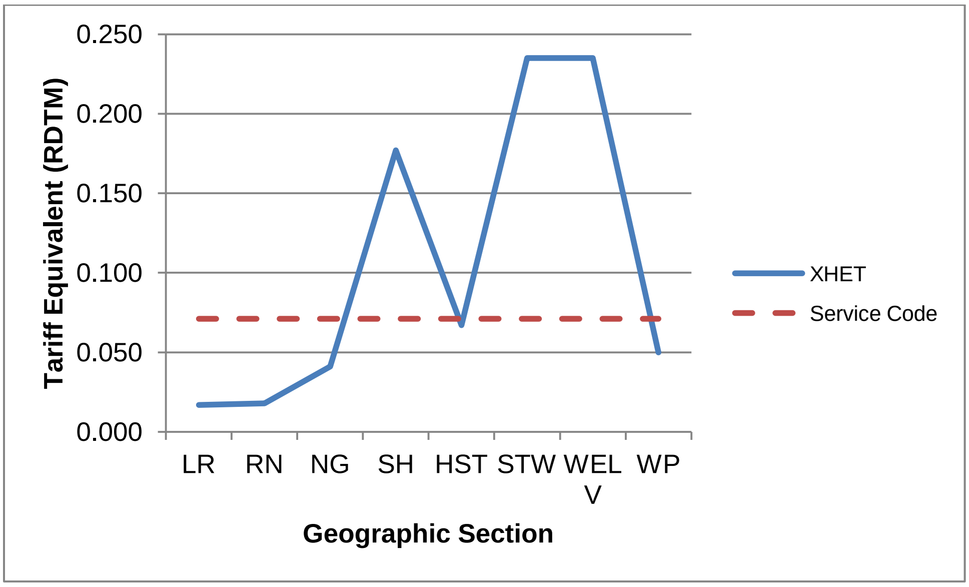


**Figure 8.18** Impact of Consolidating the Up Fast Tariff Equivalents Into a Single Service Code Charge (0800 to 0900 Hours).

The conclusion discussed in Chapter Seven that sectional rather than area based capacity utilisation measures were more effective, suggests that there does need to be some level of disaggregation of the tariffs. Figure 8.19 illustrates the concept of a split tariff that might achieve the necessary balance for Freight traffic between ease of implementation and effective incentive. The tariffs for the individual areas represent weighted averages of all the sections that form that particular area. The peaks and troughs in congestion within each area are therefore accounted for as is the relationship between each geographic section.

The additional line shows the two different Tariff Equivalents calculated for the Newark and Welwyn areas. It can be seen that the Welwyn Tariff Equivalent is substantially higher reflecting the greater capacity utilisation and therefore marginal cost of congestion in the Welwyn area. The result is a Newark area Tariff Equivalent of 0.025 minutes per train mile and a Welwyn area Tariff Equivalent of 0.138 minutes per train mile. This is compared with a combined Service Code Tariff Equivalent of 0.071 minutes per train mile.

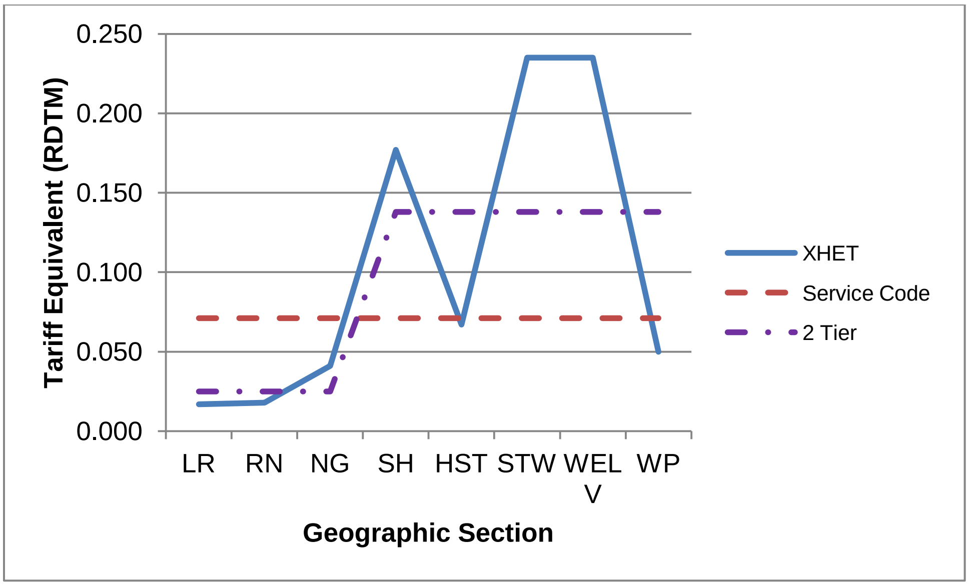


**Figure 8.19** Example Two-Tier Tariff Equivalent for the Up Fast (0800-0900 Hours).

With this approach different tariffs would apply between major nodes. For example, the Newark area might lie between the Doncaster and Peterborough nodes and the Welwyn area might lie between the Peterborough and Kings Cross nodes. Even if the tariffs remained hourly there would still be a substantial reduction in the number of geographic cells. In the sample ECML network used in this analysis there are eight geographic cells in each direction once the fast and slow line Tariff Equivalents have been consolidated. With this proposal there would be only two geographic cells in each direction.

As shown in Figure 8.19  the Welwyn area with its greater congestion would have a higher tariff than the Newark area; the large number of freight trains that pass through the Newark area but not the Welwyn area would therefore make a considerable saving. Since freight operators (like Open Access operators) have to bear the cost of the Capacity Charge themselves, lower costs would give a commercial advantage. For example, these savings could be passed onto customers.

The creation of tariffs at a more detailed level than Service Codes is of course a contentious issue given Network Rail's view that this would be too difficult and expensive to implement (Arup, 2013). However, this could be circumvented through the introduction of more Service Codes. As noted in a previous Chapter, there is a considerable amount of difference in the number of Service Codes that different franchised passenger operators have. This suggests that there is scope for increasing the number of Service Codes to enable the split tariff approach which has been suggested here.

### 8.5.4 Division of HET Based Tariffs by Time Band

One of the key tariff patterns that have emerged in the discussion of possible tariffs is the ‘smoothing’ effect of moving from an hourly Tariff Equivalent to a three-hourly Tariff Equivalent. This will obviously reduce the impact of any Congestion Charge.

One possible approach is to use split time periods in a similar manner to the split geographic tariffs suggested in the previous section. The concept behind this approach is to have hourly tariffs during the peak periods, in order to provide greater definition, but consolidated periods during the rest of the time. A possible division of time periods is shown in Table 8.5 and would apply to the Welwyn area.

**Table 8.5** Suggested Time Period Division for the Welwyn Area.

| Time Period | Hours | Type |
|---|---|---|
| 1 | Pre 0700 | Off-Peak |
| 2 | 0700-0800 | Peak |
| 3 | 0800-0900 | Peak |
| 4 | 0900-1000 | Peak |
| 5 | 1000-1300 | Inter-Peak |
| 6 | 1300-1600 | Inter-Peak |
| 7 | 1600-1700 | Peak |
| 8 | 1700-1800 | Peak |
| 9 | 1800-1900 | Peak |
| 10 | Post 1900 | Off-Peak |

Application of the time periods in Table 8.5 would therefore reduce the time band tariffs for the sample area from 16 to 10. In combination with the reduction in the geographic bands described in the previous section from 24 to 4 this would reduce the total number of tariff cells from 256 to 40.

Figure 8.20 illustrates the difference between a single time period tariff; one based on three-hourly time periods and one using the time period divisions suggested in Table 8.5. It can be seen that the main difference is in the Tariff Equivalents during the morning period. This is unsurprising given the fact that the direction illustrated (the ‘Up’) covers the flow into London. The use of a split tariff introduces greater definition during the time periods where there is the heaviest congestion. The avoidance of the ‘smoothing’ effect seen with the three-hour time periods (and even more so with a single tariff) for the whole day ensures a greater incentive to plan services outside the most congested time periods.

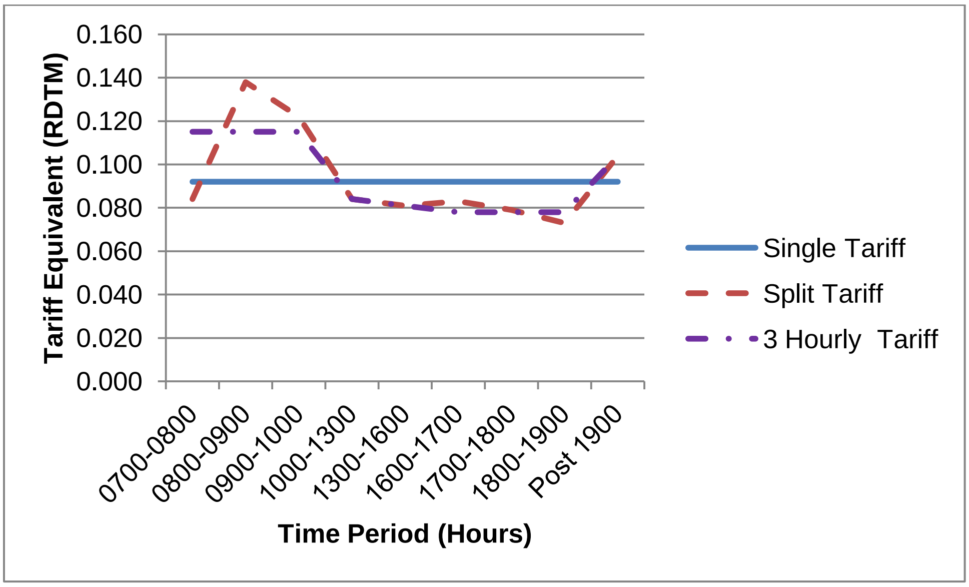


**Figure 8.20** Illustration of Three Different Types of Tariff Equivalent for the Welwyn ‘Up’ Area.

However, one issue that needs to be discussed is the potential problem of time period boundaries. The issue of a bid for a service close to a time period boundary being flexed by Network Rail into a more expensive band was specifically raised by one of the consultee responses (AECOM, 2012) to the Capacity Charge recalibration and again within the final report (Arup, 2013). This was put forward as a key reason why there should not be time period boundaries in the final capacity charge and instead it should be Service Code based (e.g. an East Coast Leeds to London train arriving at 1400 hours should pay the same tariff as one arriving at 0845).

A more effective solution, in terms of maintaining price signals, is however for each Operator to state in their bid the time period that the train belonged to. This, with the exception of entirely new trains, would reflect the time period stated in the track access contract. Network Rail would then attempt to plan that train within the stated time period. However, if following flexing to a level permitted by the contract, the train was finalised in a different price band the operator would pay the cheaper of the two tariffs. This would incentivise Network Rail to reduce congestion during the train planning process since they would be encouraged not to flex trains into the peak periods, as otherwise they would receive a lower rate of compensation for an increased risk of reactionary delay.

### 8.5.5 Use of Service Codes in the Congestion Charge Process

As discussed previously  the recalibration of the Capacity Charge in 2013 based the resulting tariffs on Service Codes due to the problems with implementation that would arise from any greater disaggregation (Arup, 2013). However, as demonstrated in the previous sections, Service Code tariffs lead to a considerable 'smoothing' of the price signals. The use of Service Codes therefore sacrifices 'incentive' in favour of 'practicality'.

The answer to this problem however, seems to be fairly straight forward. The creation of more Service Codes would produce the necessary variation in incentive whilst maintaining the existing billing system. The likely shape of such an approach has been discussed in the previous sections (e.g. the use of the split time bands) . In terms of actual implementation, this would depend on the nature of the traffic.

For passenger traffic on the ECML sample network used in this thesis, the division into additional Service Codes could be based on arrival at and departure from London Kings Cross Station. These times would determine which Service Code each particular train would be allocated to. For example, rather than having a very limited number of Service Codes those for East Coast trains could be expanded to include the different time bands shown in Table 8.5. Further sub-divisions would be necessary to separate the Leeds traffic from the Anglo-Scottish traffic for example.

Freight traffic would obviously be more complicated due to the variety of origins, routes and destinations. However, one approach that could be used to reduce this would be to base the tariffs around the core part of the route. This would be the most congested part of the train journey and likely to have the biggest interaction with other types of traffic. The tariffs for the different origins and/or destinations could then be averaged. Such an approach is illustrated in Figure 8.21.

The figure shows the amalgamation of various non-core origins and destinations to produce two separate Service Codes. It can be seen that this approach would substantially reduce the number of different groups whilst maintaining separate tariffs for the core parts of the route.

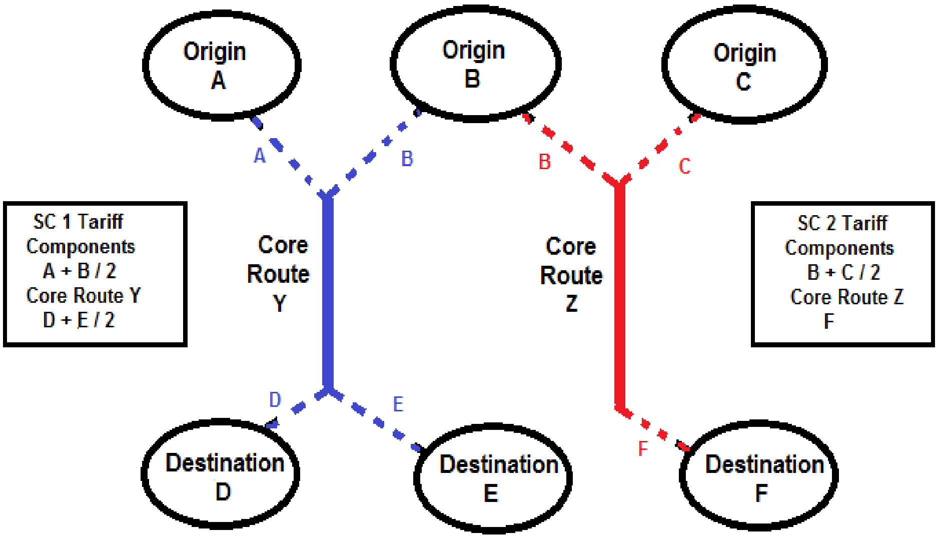

Origin A
Origin B
Origin C
A
B
B
C
SC 1 Tariff Components A + B / 2 Core Route Y D + E / 2
Core Route Y
Core Route Z
SC 2 Tariff Components B + C / 2 Core Route Z F
D
E
F
Destination D
Destination E
Destination F

**Figure 8.21** Illustration of Process to Create New Freight Service Codes as the Basis for Congestion Charging.

The actual Service Code that an operator's service belonged to would be detailed in their track access contract and in their bid for each timetable. As noted previously, Network Rail would be encouraged not to 'flex' trains into higher Service Code bands by only receiving compensation based on the documented one. New traffic would be the subject of agreement as would any dispute about existing traffic.

Although, the implementation of a new approach could be fairly complex and costly; the price signals would be much more appropriate. Network Rail would receive compensation better reflecting the actual marginal cost of congestion. As noted earlier in this thesis, the income from the Capacity Charge regime and the cost of reactionary delays are both substantial. Although, it is important to introduce a regime that is practical it is also important to introduce one that better reflects the true cost of congestion.

### 8.5.6 Charging All Traffic Versus Charging New Traffic

During the consultation process for the recalibration of the Capacity Charge for the British Rail network there was some discussion about whether a congestion charge should be levied on all traffic, as proposed, or as some respondees argued solely on new traffic (Arup, 2013).

The latter appears to have some basis in logic. Since the Capacity Charge tariff is calculated using ‘one additional train’ then it may seem reasonable to suggest that only additional trains should attract the charge.

However, selective charging potentially produces an issue about what exactly constitutes additional traffic. As described earlier in this thesis, there are a number of steps in the creation of a timetable that operates on any given day. Services can be planned in the Permanent Timetable that is finalised a number of months in advance. They can also be planned at very short term notice. Services can also be already described in existing track access contracts or contained in new ones. This variation in when trains are planned and how they are legally defined makes it difficult to be precise about what would actually constitute an ‘additional’ train. In any case all trains on the network contribute to the level of congestion.

There is also an important reason why a congestion charge should apply to all traffic. If a charge is not applied to all traffic within a time period the concept that the objective of the charge is to better use capacity within the timetable as a whole will be undermined. Only Operators introducing new services would be affected by a selective charge and only within a particular time period and for the relevant geographic sections. The ‘better use of capacity’ argument applies to all cells and not just to ones that have new traffic. For this reason Network Rail’s approach does appear to be reasonable and logical.

Finally, the idea of applying a tariff to all traffic is also consistent with the theoretical approach to congestion charging that suggests everyone should share the marginal cost of congestion equally. On a theoretical basis the charging of all traffic would therefore ensure that all traffic was subject to the appropriate price signals.

### 8.5.7 Linking the Capacity Charge with Congested Infrastructure Declarations.

As described earlier in this thesis (p9) a formal declaration of Congested Infrastructure is required to be made by Network Rail when requests for access cannot be satisfactorily met. Although, this describes a state of scarcity rather than congestion it is appropriate at this stage to consider whether it might be appropriate to combine elements of the two approaches. As noted following the declaration of Congested Infrastructure, Network Rail is required to undertake a capacity study to identify the extent of the issues. This is then followed by the production of a plan to address the shortfall in capacity. Linking declarations with the Capacity Charge could provide an incentive to Network Rail to declare all route sections with capacity issues as Congested Infrastructure. A suitable link could take the form of any section with a calculated capacity value exceeding a certain level (for example the UIC recommended values shown in Table 3.4) requiring a capacity study or sections could be ranked according to ‘capacity charge per mile’. This would encourage Network Rail to understand the underlying capacity issues. To incentivise this, Network Rail could be required to give up some of the Capacity Charge for the section if such a capacity study and improvement plan was not forthcoming (which could be used to help fund an independent study).

However, given the known congestion issues on the network referred to elsewhere in this thesis, this could lead to the production of many more capacity studies and improvement plans than the two currently produced. It is likely in the short term at least that Network Rail would be ‘swamped’ with the requirement to investigate capacity issues. The approach is therefore probably impracticable. There is also the danger that through investigating and attempting to address capacity issues on the network in this way; the problem will just be transferred elsewhere. As noted previously, the British rail network is highly interconnected.

## 8.6 Summary

This chapter has described the application of the regression results described in Chapter Seven to the creation of potential congestion Tariff Equivalents. A number of key themes have been discussed. These are:-

- The calculation of a congestion Tariff Equivalent using the concept of an ‘additional’ train.
- The methodologies used to calculate sample Tariff Equivalents for both CUI and HET based capacity utilisation measures.
- The characteristics of the different Tariff Equivalents which emerge using the various capacity utilisation measures. Three important observations have been made. Firstly, the HET based Tariff equivalents are considerably greater than the CUI based Tariff Equivalents. Secondly, there is more variation between time periods with the HET based Tariff Equivalents due to the greater ‘slope’

discussed in Chapter Seven. Thirdly, there is more variation between traffic types (i.e. using the Service Codes) with the HET based Tariff Equivalents.

- The use of three-hour periods produces a ‘smoothing’ of the Tariff Equivalents compared with one-hour periods.
- Whilst the presence of infrastructure constraints has an impact on the calculated Tariff Equivalents for the Welwyn area they do not for the Newark area. This is due to the different nature of traffic between the two areas.
- Although potentially attractive, the idea of linking the Capacity Charge and the Declaration of Congested Infrastructure is not considered appropriate due to the scale of the task this implies and the possible transfer of the problem to other locations on the network.

Several important conclusions have also been reached during this Chapter:-

1. There is greater incentive using the HET based Tariff Equivalents to plan traffic at less congested times. The implication is also that the less ‘effective’ CUI based measures will undercharge for the cost of congestion.
2. Using the XHET values for the ‘A’ and ‘β’ values with evenly spaced HET percentages would provide greater encouragement to Network Rail to effectively plan capacity utilisation than the retention of the original irregular spacing. This is because the level of compensation would be lower.
3. All sections have a bearing on the final level of reactionary delay. It is therefore important that a tariff regime is in place that covers all geographic sections if the most effective price signals are to be produced.
4. However, the use of ‘split’ tariffs that cover key areas of a journey appears to be the best compromise between effective incentives and practicality.
5. The number of time periods that are used could be reduced by only having single hours to cover peak periods (and therefore produce the required fineness of definition) with off-peak and inter-peak periods covered by three-hour time bands.

6. The adoption of points 4 and 5 above would reduce the number of tariff cells in the sectional database from 256 to 40 without it is believed compromising the nature of the incentive regime too much.
7. All traffic rather than simply additional traffic should be charged for.

# Chapter Nine
# Conclusions

## 9.1 Overview

This thesis has investigated the relationship between rail capacity utilisation and timetable performance and used the findings to consider possible mechanisms for charging for access to congested infrastructure. The background to this is a growing demand for rail travel on an already crowded network with finite capacity which is expensive and time consuming to expand. The growing interest in optimising capacity utilisation makes this thesis particularly timely.

An extensive literature review into alternative capacity utilisation measures has been undertaken and the results of this have been discussed. It is clear that there are a number of different approaches and philosophies each of which has its own merits. A substantial part of this thesis has therefore been devoted to comparing several different measures. In contrast, the meaning of timetable performance, and in the context of this thesis reactionary delay, is clearly defined in Britain. This is due to the framework of the privatised railway and in particular the existence of performance regimes between Network Rail and the train operators. In addition, the ORR's role in monitoring the success of Network Rail in delivering a reliable timetable ensures that detailed records are kept. The existence of detailed timetable and performance data means that it has been possible to undertake the analysis described in this thesis. Other explanatory measures have also been explored with the aim of understanding whether capacity utilisation alone can provide an adequate explanation of timetable performance.

The relationship between capacity utilisation and reactionary delay has been investigated using standard econometric regression techniques and 'success' measures. A number of different functional forms have been tested based on previous empirical work on the subject. The results have been described in detail and their transferability to other congested rail networks discussed. The values have been used to suggest and compare possible congestion charging mechanisms taking into account previous theoretical and practical work on the subject.

This chapter reviews the findings of the analysis carried out, assesses how appropriate the suggested congestion charging mechanisms are in light of previous work on the subject, makes recommendations for appropriate future work and finally describes the contribution that this thesis makes to this important subject.

## 9.2 Background to the Regression Analysis

### 9.2.1 The Data Set

The analysis has been carried out using data for two areas of the southern portion of the East Coast Main Line (ECML) for the December 2009 to May 2010 timetable. The ECML was chosen due to its known congestion issues. The two areas were based on different types of infrastructure ‘bottlenecks’ or constraints. The timetable was chosen due to its having a reasonable mix of traffic types. The data set is believed to provide a good representative example of congested rail networks in Britain. It is believed that the data set provides a good basis for the rigorous testing of the relationship between a variety of capacity utilisation measures and reactionary delay in order to determine which of the former is the most ‘effective’.

Two levels of detail have been used in the analysis. Firstly, the data has been examined at a sectional (or meso) level with capacity utilisation being calculated between the compulsory timing points on the network. The sectional data set matches the capacity utilisation calculations to the level that the performance data was provided. The analysis can therefore be considered ‘performance led’. This is believed to be more appropriate than the micro level approach adopted for previous work for the calculation of the Capacity Charge (a current congestion charge levied on all traffic using the British rail network). In this case the performance data was allocated to geographic sections which reflected ‘Constant Traffic Sections’ i.e. even minor timing locations were used as boundaries for the links. This latter approach can be considered ‘infrastructure led’.

Secondly, data has been analysed at an area (or macro) level. This has enabled investigation into whether capacity utilisation at the key infrastructure constraints influences the overall level of reactionary delay in the surrounding area.

In total there were twenty-four sections and four areas. The data has also been divided into 16 different hourly time-bands (between 0600 hours and 2200 hours). This gives 384 sectional cells and 64 area cells for the analysis.

### 9.2.2 Capacity Utilisation and Other Measures

The literature review identified a number of suitable capacity utilisation measures for the analysis. Three basic types of measure were identified. Firstly, there were those that calculated capacity utilisation on the basis of the volume of capacity used (Traffic Intensity and the Capacity Utilisation Index). Secondly, there were those that measured the way capacity was used and in particular the size of the gap in-front of each train (Heterogeneity Measures). These measures which were described by M.J.C.M., Dekker, R. and Kroon, L.G. (2006) were improved in this thesis by the addition of a denominator allowing the percent capacity used to be calculated. This denominator was based on the relevant planning headway or margin which determines the minimum timetabled 'buffer' between successive services. Thirdly, there were those measures which were used with the area data set; that linked the capacity utilisation at a principal constraint with the reactionary delay for the entire area. These measures are based on the Theory of Constraints concept.

The majority of previous work on capacity utilisation measures has focused on link-only utilisation. The exclusion of nodal capacity utilisation, due to the added complexity it brings, is seen as a serious omission for previous work. New capacity utilisation measures were therefore produced which modified the existing CUI and Heterogeneity measures to produce Junction CUI and Junction HET respectively. Generally, the junction nodes have been included at the end of geographic sections. This is based on the rationale that traffic approaching a junction would be most likely to suffer reactionary delays due to congestion in advance of it. The exceptions to this rule, most notably the links which include Newark Flat Crossing, have been explained in the relevant parts of the text. The inclusion of capacity utilisation measures which take into account node and link capacity utilisation therefore represents an improvement on previous analysis which focused on 'link only' capacity utilisation.

However, although junction capacity utilisation was included in the analysis; the impact on capacity utilisation of limited platform capacity at stations was excluded for being another potentially complicating factor. Instead, station nodes were used as the start or end points of the various geographic sections with the intervening link being considered 'exclusive' of them.

In total eight different capacity utilisation measures were tested on the sectional data set using either the first or second of the two types of approach described previously. Three capacity utilisation measures were tested using the area data set. The latter were designed to test the effectiveness of the Theory of Constraints concept. As described in the main body of the thesis this theory suggests that the capacity and success of a system is dictated by the capacity and flow through its principal constraint. Each of the capacity utilisation measures are described in Chapter Three and their equation given. They are summarised in Table 9.1.

**Table 9.1** Summary of the Capacity Utilisation Explanatory Variables Used in the Analysis

| Measure | Type | Scope |
|---|---|---|
| Intensity (I) | Link-Only Volume | Sectional |
| OCUI | Link-Only Volume | Sectional |
| XCUI | Link & Node Volume | Sectional |
| OHET | Link-Only Spacing | Sectional |
| AHET | Link-Only Spacing | Sectional |
| XHET | Link & Node Spacing | Sectional |
| VHETB | Link-Only Spacing | Sectional |
| VHETF | Link-Only Spacing | Sectional |
| LCUI | Link-Only Volume | Area |
| LHET | Link-Only Spacing | Area |
| EHET | Link-Only Spacing | Area |

A number of alternative measures were also tested to establish if these could complement the capacity utilisation measures or indeed replace them as effective explanatory variables for reactionary delay. Once again these were developed following a literature review. These were:-

- Timetable Complexity
- Average Distance Travelled
- Average Transit Time Variation
- Stability
- Average Entry Lateness

In addition in response to concerns that the ‘poor fit’ of the data in the recalibration of the Capacity Charge was due to bias from network effects, three additional explanatory variables were tested:-

- (Capacity Utilisation of) Time Period Before
- (Capacity Utilisation of) Section Before
- (Capacity Utilisation of) Section Following

### 9.2.3 The Regression Methodology

The organisation of the regression data into panel data is the standard approach for processing data that contains both cross-sectional and time-series data. A large number of variations were examined. The various combinations of one-way and two- way models and fixed effects and random effects models were tested for each of the different capacity utilisation measures. In all there were also five functional forms (Linear, Quadratic, 2nd Order Approximation – Linear, Exponential and 2nd Order Approximation – Logarithmic). Three recognised measures of success: the adjusted R-squared value, the t-statistic and the F-test of Joint Significance were used to determine the most successful estimator.

Timetable and performance data for the analysis was supplied by Network Rail. As noted previously care was taken to check any calculations and results as closely as possible.

## 9.3 Results of the Regression Analysis

### 9.3.1 Type of Model

The in-depth regression analysis reached a number of important conclusions. Except where specifically stated, the adjusted R-squared value was used as the decision criteria.

Firstly, for a number of reasons an Exponential functional form was deemed the preferred functional form to describe the relationship between capacity utilisation and reactionary delay.  This is  clearly logical. As the network becomes more crowded it seems likely that more traffic will be susceptible to ‘knock-on’ delays and the impact of the original delays will be greatly magnified. This also confirms the findings of previous work on the subject and in particular that for the original and recalibration of the Capacity Charge.

Figure 9.1 show this Exponential function form where RDTM equals reactionary delay per train mile; *A* equals a constant that changes by geographic location (and possibly by time series); *β* is a constant value for the Capacity Utilisation Measure and *Cap* is the percentage capacity utilisation.

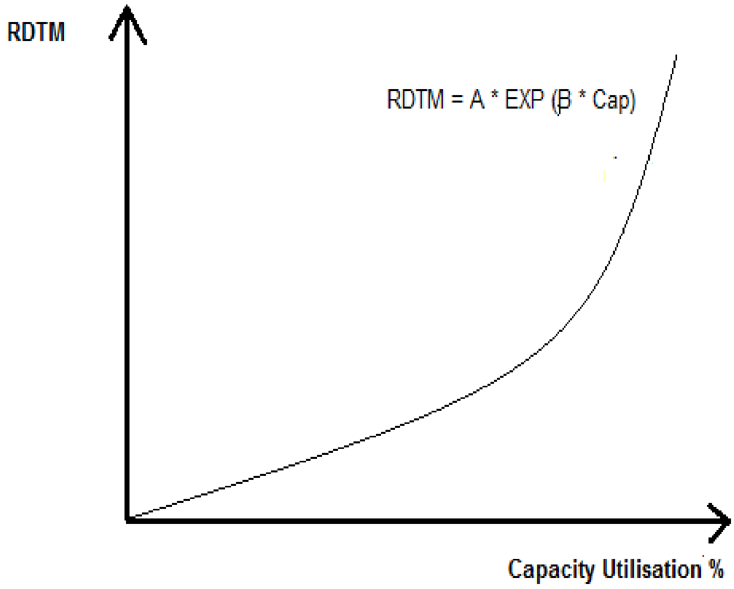


**Figure 9.1** Preferred Functional Form of the Relationship Between Capacity Utilisation and Performance (Reactionary Delay).

The exponential relationship has significant implications for the development of a tariff for congested rail networks. This is because at high levels of capacity utilisation significantly more reactionary delay will be generated than for moderate levels of capacity utilisation. Any tariffs calculated for congested parts of the rail network are therefore expected to be considerably greater than those that have more ‘spare’ capacity. A sharper price signal will therefore be sent than those that would have been developed using the linear functional form for example. This important idea will be discussed in more detail later in this chapter.

Secondly, it was concluded that a fixed effects rather than a random effects approach provided a better description of the level of reactionary delay not explained by the level of capacity utilisation. Once again this is logical and reflects the previous work on the Capacity Charge recalibration (Arup, 2013). It does seem sensible that any unexplained variation in the relationship should exhibit a fixed rather than random element, reflecting the influence of the different infrastructure that forms the geographical sections and areas in the sample network.

Thirdly, a one–way rather than a two-way model is believed to be more appropriate to test the relationship between capacity utilisation and reactionary delay. Once again this is logical. Whilst it makes sense for reactionary delay to be affected by differences in the relevant geographical section or area; it seems much less likely that variation in the time period will also be a factor other than through changes in the level of Capacity Utilisation itself.

These conclusions reflect the previous work on the Capacity Charge recalibration (Arup, 2013). In summary, therefore the regression analysis carried out as part of the research for this thesis reflects key elements of previous work on the subject and in particular the findings of the 2013 recalibration of the Capacity Charge (Arup, 2013). This previous work had also concluded that the relationship between capacity utilisation and reactionary delay was of an Exponential functional form, using a one-way model (geographic data) with fixed effects.

### 9.3.2 Sectional Capacity Utilisation Measures

Before discussing the results for the individual capacity utilisation measures, it should be noted that all of the ones included in the regression analysis were found to be significant explanatory variables using the t-statistic and F-test as appropriate. The analysis carried out for this thesis does therefore confirm the findings of previous research that capacity utilisation is a very important factor in determining reactionary delay, however the former is measured. The difference between the various capacity utilisation measures is therefore in the level of effectiveness as an estimator of reactionary delay.

One of the key conclusions from the analysis is that the effectiveness of capacity utilisation measures at predicting levels of reactionary delay is indeed improved when the former takes into account movements at junctions. The junction variants of both the CUI and HET capacity utilisation measures consistently produced better results than the associated ‘link’ only capacity utilisation variants.

A key conclusion of the work is that the HET based measures are more successful than the CUI and Intensity based capacity utilisation measures. In all cases, HET measures were found to be more effective at predicting levels of reactionary delay using the adjusted R-square value as the success measure. This is believed to be the first time that these two different types of capacity utilisation measure have been directly compared. The finding is logical given the fact that the former describes how much traffic there is whilst the HET measures are based on how capacity is actually used and in particular the size of the 'buffer' in-front of trains. As described in Chapter Three, many researchers have linked the amount of reactionary delay to the size of this 'buffer'. As noted earlier, the comparison of the different capacity utilisation measures has been made possible through the conversion of the HET based measures into percentages.

One key conclusion was that the even-spacing of traffic in a timetable would lead to reduced levels of reactionary delay compared to a bunching of trains. A number of examples from the data set used in this analysis were used to illustrate the advantage to be gained from this even-spacing. One reason HET is more successful than CUI is that the latter assumes identical trains are always evenly spaced whilst the former will give a higher percentage utilisation if there is evidence of traffic 'bunching'. Although, there is some suggestion that the selective 'bunching' or 'flighting' of trains might be a useful strategy in reducing reactionary delay (Watson, 2008); the conclusion is that in the majority of cases even-spacing is the most effective means of reducing overall reactionary delays.

As described earlier, the addition of junction moves improves the effectiveness of the HET based capacity utilisation measures. In-fact, Junction HET (or XHET) is the most effective of all the different capacity utilisation measures examined at predicting levels of reactionary delay. Apart from the basic link- only HET measure (or OHET) three other HET based measures have been used to explore the relationship between capacity utilisation on geographic sections and reactionary delay. Two of these are intended to take into account the possible greater impact of 'vulnerable' trains on reactionary delay totals. The third measure examines the theory that the gap at the end of a section, or the 'arrival' buffer between trains, rather than the minimum gap is the important one to measure.

The idea behind measuring vulnerable trains was to investigate Carey's suggestion (1999) that although the gaps between trains were important in determining the likely level of reactionary delay generated by a timetable some gaps were more important than others. In this thesis, vulnerable trains have been taken to mean those trains that are not operated by franchised passenger operators which form the bulk of the services found in the sample timetable. 'Vulnerable trains' are therefore those services operated by one of the two open access companies (Hull Trains and Grand Central Trains) or one of the freight operators. In the sample timetable these are often 'one-off' trains which have to 'fit in with' the franchised passenger operators more frequent services. The two ‘vulnerable’ measures either assume that the ‘one-off’ train itself has a greater risk of reactionary delay or the train following it does (VHETB and VHETF respectively). The two measures were calculated by putting greater weighting on the vulnerable gaps. However, although both variables are significant the results show they are similar estimators of reactionary delay to the basic OHET, with VHETB ‘performing’ better of the two. It is however recognised that the approach adopted was rather crude, the weighting of ‘vulnerable’ gaps being simply twice that of other ‘non-vulnerable’ services.

The use of the ‘arrival’ gap between trains was found to be the least successful of the HET capacity utilisation measures. This is somewhat contrary to Vromans, M.C.J.M., Dekker,R. and Kroon, L.G.’s expectations (2006). However, once again this finding is logical as it means that the minimum ‘buffer’ time between trains is the most important determinant of reactionary delay wherever that may occur in the section. As discussed earlier in the thesis whilst the concept of the ‘arrival gap’ recognises that the risk of reactionary delay increases as a ‘fast’ train catches a ‘slow’ train at the end of a section; it does not recognise the other possibility that a ‘slow’ train at the start of a section may be delayed by a preceding ‘fast’ train.

### 9.3.3 Other Sectional Variables

The investigation of the effectiveness of ‘other’ explanatory variables using the t-statistic revealed that a number of these non-capacity utilisation measures became significant when included in the regression specification. Interestingly though there was not a great deal of consistency, some ‘other’ variables became significant with certain capacity utilisation variables but not with others.

The complexity of the timetable (as measured by the number of Service Codes) was consistently identified as a good complimentary explanatory variables. The success of the timetable complexity variable is not surprising given the fact that it adds an additional element to the equation. It is logical that timetables with greatly increased complexity due to a large number of different types of traffic will be at greater risk of reactionary delay than those with low complexity. It also supports the previous research described in Chapter Three of this thesis.

Interestingly, the ‘Section Following’ variable was only found to be significant with XHET, further increasing the adjusted R-squared value. Given that generally the junction nodes were included at the end of links, this suggests the inclusion of the SFCAP variable helps explain the influence of capacity utilisation following the node on levels of reactionary delay.

Timetable Complexity was also found to be a significant explanatory variables when used on its own. Surprisingly, the variable was found to be more significant than either of the CUI based variables or the Intensity variable. One of the ‘criticisms’ of these ‘volume of capacity used’ variables is that they do not give an insight into how that capacity is actually used in a timetable. This finding supports the view that this is an important factor in the estimation of reactionary delay.

The overall conclusion was that other variables could, when added to capacity utilisation variables, improve the explanation of the cause of reactionary delay. Examining the level of correlation did however show that in some cases there was a substantial degree of overlap between the explanatory variables. From the perspective of considering the analysis required to transfer these relationships to a national level; it was felt that the additional explanatory power provided by these ‘other’ explanatory variables was not sufficient to justify their inclusion in the specification. A parsimonious relationship is also consistent with the approach adopted for the recalibration of the Capacity Charge (Arup,2013).

### 9.3.4 Area Capacity Utilisation Variables

One surprising outcome of the research described in this thesis is the limited success of the explanatory variables developed to test whether the Theory of Constraints could be used to describe the relationship between capacity utilisation and reactionary delay. Although the two explanatory variables developed to test the theory (LHET and LCUI) were significant they were found to be less effective than EHET. This measured the smallest gap in an area wherever that might occur.

The ‘success’ of EHET reinforces the findings with the sectional capacity utilisation measures that the minimum gap between trains wherever that might occur is a key explanatory factor in the development of reactionary delay. Although logically the Theory of Constraints sounds an attractive concept in describing the relationship between capacity utilisation and timetable performance; capacity utilisation at locations other than the primary constraint were found in the analysis to have a significant influence on the observed level of reactionary delay.

It was also found that a Linear functional form best describe the relationship between the area capacity utilisation measures and reactionary delay. This is not believed to be intuitive as it implies a static increase in the rate of reactionary delay as the network becomes more crowded. It was suggested that this finding might be due to the small size of the data set.

For a number of reasons it was concluded that the sectional capacity utilisation variables provided a more accurate explanation of the occurrence of reactionary delay than the area variables. In other words, explanatory variables at a meso rather than a macro scale provided a more appropriate fit for the data.

### 9.3.5 Overall Summary

An explanatory variable based on measurement of the smallest gaps at links and nodes (i.e. XHET) within an exponential relationship using a one-way model with ‘fixed effects’ therefore provides the preferred prediction of the reactionary delay observed in the data set used for this analysis.

The reason for the success of the HET based measures, and one of the key conclusions of this thesis is that the level of ‘bunching’ of traffic is a critical factor in determining the overall amount of reactionary delay that is generated. Whilst the HET based measures take into account heterogeneity in both identical and non-identical traffic, CUI measures only take into account the latter. This helps explain the greater success of the HET based measures.

## 9.4 The Development of a Congestion Charge

### 9.4.1 Overview

As described in the previous paragraphs, the first part of this thesis was concerned with identifying the most effective way of modelling the relationship between capacity utilisation and timetable performance (reactionary delay). The regression analysis established that a measure developed from the Heterogeneity (or as described in this thesis ‘HET’) approach proposed by Vromans, M.C.J.M, Dekker, R. and Kroon, L.G. (2006) was a more successful estimator than the CUI approach, the standard method of capacity utilisation measurement in Britain. The next step was to consider the implications of these findings for the charging of congested rail networks.

Example Tariff Equivalents were produced for the sample rail network using a similar methodology to the one used for the calculation of Britain’s Capacity Charge tariffs. A key aspect of the approach is the calculation of the cost of the reactionary delay generated by one ‘additional’ train. These calculations used the constant values obtained during the regression analysis. The use of the marginal cost to calculate Tariff Equivalents is logical and supported by theory.

### 9.4.2 Service Code Tariff Equivalents

The first part of the calculations used combined Tariff Equivalents to produce single high-level Tariff Equivalents divided solely by Service Code. This replicated the Tariff Equivalents produced for Britain’s Capacity Charge. The second part of the Tariff Equivalent calculations considered alternative approaches and this included considering charges based on different time bands and geographic sections. The methodology used to calculate both types of Tariff Equivalent is described in Chapters Five and Eight.

One key finding was that the HET regression values produced substantially higher Tariff Equivalents than the CUI regression values. Given that the HET measures were found to be more effective explanatory variables for reactionary delay, this suggests that the true cost of congestion is nearer to that predicted by the HET approach than the CUI approach. In other words, CUI under-estimates congestion costs.

Another important conclusion was that the division of Tariff Equivalents into time periods produces a clear pattern. Peak traffic periods are accompanied by peak Tariff Equivalents and off-peak (or less congested times) are accompanied by lower Tariff Equivalents. The move from hourly tariff bands to three hourly tariff bands does however produce a ‘smoothing’ effect. Once again this is logical and shows that the move to Service Code based Tariff Equivalents considerably lowers any incentive that traffic might have to operate at less congested times.

In terms of the division of Tariff Equivalents into geographic sections it was concluded that retention of separate Tariff Equivalents for these sections produced clear patterns. Interestingly, those for the Welwyn Area could be seen to reflect the presence of Welwyn Viaduct and to a lesser extent Hitchin Junction constraints. In contrast, the highest Newark Area Tariff Equivalents did not always correspond to the sections associated with the physical infrastructure constraint at Newark Flat Crossing (Newark to Retford and vice versa). Instead the highest Tariff Equivalents were generally found in the Grantham to Newark sections. This reflects the fact that the primary constraint in the Newark area arose due to the mix of traffic in the timetable.

Once again though, the combining of Tariff Equivalents into single ones based on Service Codes leads to a ‘smoothing’ effect. This is less important than the consolidation of time bands unless the traffic in question (e.g. freight traffic) potentially has alternative routes with different tariffs that could be used i.e. an incentive exists. However, East Coast Trains’ paths for example pass through each of the eight consolidated sections in the sample network. It is therefore immaterial that the Grantham to Newark section is more expensive than the Newark to Retford section as both are passed through.

A key conclusion is that although some differential is provided by Service Codes, there is a lack of significant incentive between individual tariffs. Train Operators and Network Rail are not encouraged to seek to plan services at less congested times (in the case of all operators) or on less congested routes (in the case of Freight operators who might have some choice). Instead the Capacity Charge operates as a compensatory mechanism for Network Rail to seek to recover the marginal cost from increased reactionary delay generated by the growing demand for train paths on the British Rail network.

### 9.4.3 Alternative Thoughts on the Calculation of a Congestion Charge

A number of alternative thoughts on how tariffs could be levied were investigated.

One conclusion was that although the HET based Tariff Equivalents were more realistic congestion charges, these should not be used unmodified. As noted above one of the problems of the current Capacity Charge is that rather than providing an incentive to the parties involved to improve capacity utilisation they simply give compensation to Network Rail. Adopting the HET Tariff Equivalents which reflect the timetable used as the basis for the calculations would, as noted, lead to larger amounts of money being paid to Network Rail than if the CUI Tariff Equivalents had been adopted. The tariff would therefore reflect the level of ‘bunching’ in the base timetable.

Theoretically, Network Rail could reduce reactionary delay by recasting the timetable to reduce bunching but still receive a high level of compensation. They would therefore be handsomely rewarded for fulfilling one of their licence obligations which is the efficient use of capacity. By basing the tariffs on evenly spaced capacity utilisation, Network Rail would receive a lower amount of compensation (since the calculated capacity utilisation would itself be lower) but still be optimised to reduce reactionary delay through improvements to the timetable.

A second important conclusion was that all trains should be charged. This is because every train contributes to the level of congestion and hence to the risk of reactionary delay. In effect therefore every train can be considered the ‘marginal’ train. This is in line with current theory. Another reason is that unless every train is charged then Operators will not be incentivised to work with Network Rail to optimise capacity use within a timetable.

A third important conclusion was that the principle of charging for non-congested locations and routes should be very carefully considered. Currently the Capacity Charge levies a tariff for all parts of the network. However, returning to the previous regime of ‘de-minimis’, i.e. not charging the smallest tariffs, may further encourage traffic (principally freight) which is able to switch routes to do so. This could help achieve a more efficient use of traffic on the network. However, the danger is that without the balance of tariffs those diversionary routes may themselves start to become over-loaded with traffic. Furthermore, to encourage switching of traffic the waiving of very small tariffs may be insufficient. As discussed on Page 87 of this thesis, Arup found that the impact of the de-minimis threshold on tariffs was marginal.  On balance therefore it appears that the charging for all routes is the most sensible way forward.

Finally, it has been concluded that there needs to be sufficient differentiation between geographic and time-series tariffs if an adequate incentive to optimise the use of capacity is to be created. It is suggested that different tariffs for routes between major nodes are created and tariffs divided into time-bands are retained. As noted in Chapter Eight it is believed that this could be achieved through the creation of more Service Codes.

## 9.5 Recommendations for Further Work

Although this thesis has covered a great degree of material, there are some issues that would profit from further work. These can be divided into firstly, aspects that were deliberately excluded from the analysis and secondly, into areas that the results of the research suggest could benefit from additional work.

The recommendations for further work are as follows:-

- The conclusions are based on a sample network of two portions of one of Britain’s mainlines. Although, it is firmly believed that the findings are transferrable it would be useful to apply the techniques developed for this thesis to other routes.

- A more comprehensive analysis of junction capacity utilisation would be helpful. The work carried out for the thesis only examined traffic flow from the perspective of the ECML. It would be profitable to expand the work to include all ‘links’ adjacent to a junction.
- It was generally assumed that junction capacity utilisation occurred at the end of a link. Further investigation into the most appropriate ‘position’ of a node in a network would be useful.
- The weighting between junction and link spacing for XHET was an arbitrary 50:50. This weighting therefore assumes that the two different types of capacity utilisation have the same impact on performance. Future work could test various different weights.
- It is believed that the effectiveness of the derived relationship between capacity utilisation and performance could be further improved if station capacity utilisation was successfully included in the analysis.
- The weighting for the gaps before and after Vulnerable trains (i.e. the variables VHETB and VHETF) was also set at an arbitrary double that of other trains. Although, producing similar results to OHET, the concept does appear to be theoretically sensible. It would therefore profit from investigating different weights. Secondly, experiments with different trains classed as ‘Vulnerable’ could be carried out. For the thesis these were taken to be Freight and Open Access trains. One possibility might be to include empty coaching stock (ecs) services.

## 9.6 Contribution of this Thesis

This thesis has explored the relationship between capacity utilisation and timetable performance on congested rail networks through the comparison of different measures not previously looked at together. The heterogeneity measure suggested by Vromans, M.C.J.M., Dekker, R. and Kroon, L.G. (2006) in relation to Dutch rail networks has following the introduction of a denominator (allowing the calculation of a percentage capacity utilisation) been found to be more ‘successful’ for the sample network used than the current standard capacity utilisation measure used in Britain (CUI). The findings support the belief that how capacity is used is a more important determinant of levels of reactionary delay than simply how much capacity is used.

The work has concluded that the an Exponential form is the  preferred functional form for describing the relationship between capacity utilisation and reactionary and this supports previous research on the matter. A one-way, fixed effects model has been found to be the most appropriate approach and once again this supports previous work.

The thesis has concluded that measures taking into account junction and link capacity utilisation, which have been developed as part of this work, are more effective explanatory variables than those considering link-only capacity utilisation. This confirms the belief that nodal capacity utilisation is a key factor in the development of reactionary delay.

A number of ‘other’ explanatory variables have been considered but a parsimonious specification with only capacity utilisation variables is preferred. This thesis concludes that capacity utilisation is a prime factor in determining the level of reactionary delay, although it is important to take into account local differences in the infrastructure (this being accounted for by a one-way fixed effects model).

Although, superficially attractive the Theory of Constraints has not been found to be a particularly useful concept in explaining the relationship between capacity utilisation and congestion. Instead this thesis demonstrates that both infrastructure and timetable constraints can act in unison to determine overall levels of reactionary delay. This thesis has demonstrated that analysis at a meso level is more effective than at a macro level. This is because the former is better at taking local factors into account.

This thesis has used these findings to examine possible charges for congested rail networks in Britain. It concludes that all trains and probably all routes should be charged for. This is in line with economic theory and the conclusions reached during the recalibration of Britain’s Capacity Charge. However, in order to create the correct price incentives it is believed that there needs to be greater differentiation of tariffs between time bands and geographic location than those achieved using the current Service Code system. It is felt this could simply be achieved in Britain through the creation of more Service Codes.

Since passenger operators are able to charge a price differential on tickets for peak and off-peak travel and for travelling via different routes there seems to be no apparent reason why this cannot be possible with Britain’s

Capacity Charge. Without this greater differentiation it is believed that Britain’s Capacity Charge is simply a compensation mechanism for Network Rail rather than an effective congestion charge.

## List of References

Abril, M., Barber, F., Ingolotti, L., Salido, M. A., Tormos, P. and Lova, A. 2008. An Assessment of Railway Capacity *Transportation Research Part E* ***44****(5)* pp 774-806.

AECOM. 2012. Technical Note. Network Rail Consultation Response – Capacity Charge Prepared by S.Shapiro. [Accessed 17 October 2012] Available from: http://tinyurl.com/ptx3652

Alliance Rail Holdings Ltd. 2012. Periodic Review 2013 – Consultation on the Capacity Charge. Response from Alliance Rail Holdings. Letter from C.Hanks [Accessed 17 October 2012]. Available from: http://tinyurl.com/nl6jux3

Alliance Rail Holdings Ltd. 2013. Network Rail's Strategic Business Plan - Alliance Rail Holdings Limited (Alliance) Response Letter to ORR dated 19/2/13.. [Accessed 04 September 2013]. Available from: http://tinyurl.com/qyfpzju

Armstrong, J., Blainey, S., Preston, J. and Hood, I. 2011. Developing a CUI based approach to Network Capacity Assessment *4th International Seminar on Railway Operations Modelling and Analysis. 16/18 February 2011, Rome: University of Rome* [no publisher] [no place] pp 16-18.

Armstrong, J. Preston, J., Potts, C., Bektas, T. and Paraskevopolus, D. 2013. Developing Capacity Utilisation Methods and Limits for Railway Nodes*. 5th International Seminar on Railway Operations Modelling and Analysis, 13 /15 May 2013, Copenhagen. Technical University of Denmark,* [no publisher] [no place] pp 42-54

Arup, 2013. *Recalibrating the Capacity Charge for CP5. Final Report. Issue 2, 24 May 2013*. [Online] London, UK: Network Rail [Accessed 05 June 2013]. Available from: http://tinyurl.com/ohs675f

BBC News. 2013. Shrewsbury to London Rail Plans Rejected Again [Online]. 31 July 2013. [Accessed 27 March 2014].Available from: http://tinyurl.com/l95c7pn

Brunel, J. Marlot, G. and Perez, M. 2013. Measuring Congestion in Rail Sector: The French Experience *13th WCTR July 13-15 2013. Rio de Janiero. Brasil*. [no publisher] [no place] [no pagination]

Burdett, R. L and Kozan E. 2006 Techniques for absolute Capacity Determination in Railways *Transportation Research Part B: Methodological* **40**(8) pp616-632.

Business Infrastructure Commission. 2011 Tackling the Infrastructure Puzzle. [Online] London. British Chamber of Commerce. [Accessed 01 August 2013] Available from: http://tinyurl.com/kp3b8ob

Butcher, L. House of Commons Library. 2010. SN/BT/364 *Railways; West Coast Main Line.* [Online] London, UK; House of Commons Library. [Accessed -1 August 2013] Available from: http://tinyurl.com/lpoxpsb

Button, K. 2004. The Rationale for Road Pricing: Standard Theory and Latest Advances .In Santos G. (ed) *Road Pricing : Theory and Evidence Research in Transportation Economics Volume 9*. Elsevier, Oxford UK pp3-26.

Carey. M. 1999 *Ex* Ante Heuristic Measures of Schedule Reliability*. Transportation Research Part B; Methodological* **33** (7) pp473-494.

Carey. M. and Kwiecinski. A.1994. Stochastic Approximation to the Effects of Headways on Knock-On Delays of Trains *Transportation Research B* **28B** (4) pp251-267.

Centro. 2012 Consultation on the Capacity Charge. Centro Response. Downloaded from Network Rail website [Accessed 17 October 2012] Available from: http://tinyurl.com/lzc2uvk

DB Schenker Rail (UK) Ltd. 2012. PR13 Consultation on the Capacity Charge. Letter from N.Oatway to Network Rail. [Accessed 17 October 2012] Available from: http://tinyurl.com/k5byrkz

De Palma, A. and Lindsey, R. 2011 Traffic Congestion Pricing Methodologies and Techniques*. Transportation Research Part C* **19 (6)** pp 1377-1399.

Delay Attribution Board. 2011 *Delay Attribution Guide Issue Dated 1st April 2011*. [Online] London, UK: Delay Attribution Board. [Accessed 05 August 2011] Available from: http://tinyurl.com/q9eohue

Department for Transport, 2007a *Delivering a Sustainable Railway* [Online] London, UK: National Archives [ Accessed 05 September 2013] Available from: http://tinyurl.com/nxjqce8

Department for Transport. 2007b *Rail Technical Strategy July 2007.* [Online]. London, UK ; National Archives.[Accessed 05 September2013] Available from: http://tinyurl.com/kk9joy9

Department for Transport. 2008. *Network Modelling Framework (NMF) and Appraisal for HLOS – The Evidence Pack.* [Online] London, UK: National Archives. [Accessed 15 August 2013] Available from: http://tinyurl.com/ncjxp6h

Department for Transport. 2011 *Implementation Plans for technical specifications for interoperability. Policy : Expanding and Improving the Rail Network. Document : ERTMS National Implementation Plan* [Online] London, UK: Department for Transport.. [Accessed 05 September2013] Available from: http://tinyurl.com/bhl92ls

Department for Transport. 2012a *Railways Act 2005 Statement for CP5* [Online] London, UK: Department for Transport. [Accessed 02 September 2013] Available from: http://tinyurl.com/pke5oom

Department for Transport. 2012b *Review of the Government's Strategy for a National High Speed Rail Network* [Online] London, UK; Department for Transport. [Accessed 15 August 2013] Available from: http://tinyurl.com/m78wmml

Dingler, M.H., Lai, Y. and Barkan, C.P.L. 2009. Impact of Train Type Heterogeneity on Single-Track Railway Capacity. *Transportation Research Record : Journal of the Transportation Research Board*,**2117**, Transportation Research Board of the National Academies, Washington D.C. 2009 pp41-49

Dougherty , C. 2011 *Introduction to Econometrics*. 4th ed.. Oxford. UK: Oxford University Press.

European Conference of Ministers of Transport (ECMT). 2005 *Railway Reform and Charges for the Use of Infrastructure"* Paris: OECD Publications Service.

Eddington Sir R. 2006. *The Eddington Transport Study, Main Report, Volume 1 : Transport's Role in Sustaining the UK's Productivity and Competiveness.* Norwich, UK. HMSO.

ERTMS. 2013 Increasing Infrastructure Capacity. How ERTMS improves railway performance. Fact Sheet No.10 [Accessed 06 September 2013]. Available from: http://tinyurl.com/q42fahs

Faber Maunsell. 2007 *Capacity Charge Tariff PR2008. Recalculating the Capacity Charge for PR2008 Report for Network Rail. October 2007.*

[Online] Network Rail: London, UK. [Accessed 04 September 2013] Available from: http://tinyurl.com/kt7zuzc

Ferreira. L. 1997. Rail track infrastructure ownership: Investment and operational issues. *Transportation,* **24** (2), pp183-200.

Foster Sir A. 2010 *A Review of the Intercity Express Programme* . [Online] Railway Archives : London, UK. [Accessed on 6 September 2013] Available from: http://tinyurl.com/o4nhts9

Freightliner Ltd. 2012 Periodic Review 2013 – Consultation on the Capacity Charge. July 2012. Letter from A.Johnston to Network Rail. [Accessed 17 October 2012]. Available from : http://tinyurl.com/kr3gzb7

G.B. Railfreight Ltd. 2012. G.B. Railfreight Ltd: response to periodic review 2013 – consultation on the capacity charge Letter from I.Kapur to Network Rail. [Accessed 17 October 2012]. Available from: http://tinyurl.com/lulmvw9

Gibson, S. 2003. Allocation of capacity in the rail industry *Utilities Policy **11**(1) pp 39-42*

Gibson, S., Cooper, G. and Ball, B. 2002. The evolution of capacity charges on the UK rail network. *Journal of Transport Economics and Policy* **36**(2) May 2002 pp 341-354.

Gille, A., Klemenz, M. and Siefer, T. 2010. Applying multiscaling analysis to detect capacity resources in railway networks. In: Hansen, I. ed. *Timetable Planning and Information Quality* Southampton, UK. WIT Press. , pp73-84

Goldratt E.M. and Cox J. 2004 *The Goal. A Process of Ongoing Development* 3rd ed., Aldershot, UK: Gower Publishing.

Goodwin, P. 2004 *The Economic Costs of Road Traffic Congestion*. *A discussion paper published by the Rail Freight Group*. London UK ESRC Transport Studies Unit, University College, London.

Goverde R.M.P. 2007. Railway Timetable Stability Analysis using Max-Plus System Theory. *Transportation Research Part B* **41** (2) pp179-201

Haith, J., Johnson, D. and Nash, C. 2014. The case for space: the measurement of capacity utilisation, its relationship with reactionary delay and the calculation of the Capacity Charge for the British rail network. *Transportation Planning and Technology* **37** (1) *February 2014 Special Issue: Universities' Transport Study Group UK Annual Conference 2013.*

Hammond, S. and Dormer, A. 2013 *The Guardian Professional Website* DfT and Hitachi defend the Intercity Express Programme 20 December 2013.

[Online] [Accessed 27 March 2014]. Available from: http://tinyurl.com/n8kfvs9

Higgins, A., Kozan, E. and Ferreria, L. 1995. Modelling Delay Risks Associated with Train Schedules *Transportation Planning and Technology*. **19**(2) pp89-108.

HS2 Ltd 2013a. *Guidance. High Speed Two: an engine for growth* [Online] London: Government [Accessed 20 September 2013] Available from: http://tinyurl.com/paw7rro

HS2 Ltd. 2013b. *The Strategic Case for HS2* [Online] London: Government [Accessed 16 December 2013] Available from: http://tinyurl.com/m3q8rft

Huisman, T and Boucherie R.J. 2001. Running Times on Railway sections with heterogeneous traffic *Transportation Research Part B; Methodological* **35** (3) pp271-292.

Imperial College London. 2013. The link between congestion related delay and capacity utilisation in railways. *Appendix B, in: Arup Recalibrating the Capacity Charge for CP5 Final Report*. *Issue 2, 24 May 2013.* [Online] London, UK; Network Rail [Accessed 05 June 2013]. Appendix B. Available from: http://tinyurl.com/ohs675f

Johnson, D and Nash.C. 2008. Charging for scarce rail capacity in Britain: A Case Study *Review of Network Economics March 2008* **7** (1) pp 53-76.

Kennedy, P. 2008. *A Guide to Econometrics*. 6th ed. Oxford, UK: Blackwell.

Khadem-Sameni, M., Pretson, J. and Armstrong, J. 2010 Railway Capacity Challenge: Measuring and Managing in Britain *2010 Joint Railway Conference April 27-29 2010, Urbana, IL, USA* [no publisher] [no place] pp1-8.

Khadem-Sameni, M. 2012. *Railway Track Capacity; Measuring and Managing.* Ph.D Thesis. University of Southampton, UK.

Kraft, E. R. 1982 *Jam Capacity of Single Track Rail Lines Transportation Research Forum Proceedings **23***, Washington D.C. pp 461-471.

Krueger, H. 1999. Parametric Modelling in Rail Capacity Planning*1999 Winter Simulation Conference* 5/12 to 8/12 1999, Phoenix, Arizona [no publisher] [no place] pp1194-2000

Landex, A. 2008 *Methods to estimate railway capacity and passenger delays* Ph.D Thesis, Technical University of Denmark, Department of Transport.

Lindfeldt, A. 2012 *Congested Railways. Influence of Infrastructure and Timetable Properties on Delay Propagation* Licentiate Thesis in Transport Science. School of Architecture and the Built Environment. Division of Transport and Logistics. Department of Transport and Economics. Royal Institute of Technology, Stockholm. Sweden.

Lindfeldt, O. 2010 *Railway Operation Analysis. Evaluation of Quality Infrastructure, and timetable on single and double-track lines with analytical models and estimation* Ph.D Thesis. Division of Traffic and Logistics. Department of Transport and Economics. Royal Institute of Technology. Stockholm, Sweden.

McNulty, R. 2011 *Realising the Potential of GB Rail. Report of the Rail Value for Money Study, Summary Report*. [Online] [Accessed 01 May2011] Available from: http://tinyurl.com/m523uy7

Mabin, V.J. and Balderstone, S.J. 2000. *The World of the Theory of Constraints : A Review of the International Literature*. Boca Raton, Florida, USA:  St.Lucie Press.

Mohring, H. 1970. The Peak Load Problem with increasing returns and pricing constraints *The American Economic Review*, **60**(4) pp 693-705.

MVA Consultants .2010. *Seeing Issues Clearly. West Coast Main Line Timetable Performance Modelling Lessons Learnt*. Report produced in conjunction with First Class Partnerships for Network Rail and ORR. [Online] ORR: London. UK [Accessed 17 May 2013] Available from: http://tinyurl.com/q6sl4op

Nash, C.A. 1982*. Economics of Public Transport*. London. UK; Longman.

Nash, C.A.  2005. Rail Infrastructure Charges in Europe. *Journal of Transport Economics and Policy*, **39**(3) pp259-278.

Nash,C., Coulthard, S. and Matthews, B. 2004. Railtrack charges in Great Britain – the issue of charging for capacity *Transport Policy  **11** (4) pp 315-327*

Nash, C., Johnson, D. and Tyler, J. 2006. *Scoping Study for Scarcity Charges,* Final Report for ORR. York: ITS Leeds and Passenger Transport Networks.

National Rail Enquiries. 2013. *Ticket Types* [Online] [Accessed on 20 August 2013] Available from ; http://tinyurl.com/ogx2y3o

Network Rail. 2008a. *East Coast Main Line Route Utilisation Strategy, February 2008*. [Online][Accessed 02 July 2014]. Available from: http://tinyurl.com/pmgblqd

Network Rail 2008b *List of Capacity Charge Rates: 2009/10 Prices* [Online][Accessed 09 September 2014] Available from: http://tinyurl.com/pjtjcff

Network Rail. 2009a *Route Utilisation Strategies. Technical Guide*. Dated December 2009. [Online] [Accessed 04 September 2013] Available from : http://tinyurl.com/pbmmu6z

Network Rail. 2009b *Rules of the Plan. 2009 Timetable. London North Eastern (South) Final Version. Issued 24th April 2009*. Midlands Train Planning Office, Birmingham. Network Rail.

Network Rail. 2010a. *East Coast Main Line 2016 Capacity Review, An addendum to the East Coast Main Line Route Utilisation Strategy*. [Online] [Accessed 02 July 2014]. Available from: http://tinyurl.com/obce7r7

Network Rail. 2010b *Network Rail Infrastructure Limited. Regulatory Financial Statements 2010. Year Ended March 2010*.. [Online] [Accessed 06 September 2013] Available from : http://tinyurl.com/osjodus

Network Rail. 2011a. *Network Rail Infrastructure Limited. Regulatory Financial Statements 2011. Year Ended March 2011*. [Online] [Accessed 06/09/13] Available from; http://tinyurl.com/ol9bppj

Network Rail. 2011b. *PR 13 Initial Industry Plan Supporting Document. Definition of Proposed CP5 Enhancements.* [Accessed 23 August 2013] Available from: http://tinyurl.com/klqakqb

Network Rail. 2012a. *Control Period 4 Delivery Plan Update 2012*. [Accessed 13 August 2013]. Available from: http://tinyurl.com/n9fdqrl

Network Rail. 2012b. *Network Rail Infrastructure Limited. Regulatory Financial Statements 2012. Year Ended March 2012*. . [Online] [Accessed 06 September 2013] Available from: http://tinyurl.com/p3a5ezr

Network Rail. 2012c. *Periodic Review 2013 – Consultation on the Capacity Charge*. July 2012 [Accessed 31 July 2012] Available from: http://tinyurl.com/ol7xbv8

Network Rail. 2012d. *Preliminary Conclusions on the Capacity Charge*. Document dated 26 September 2012. [Accessed 17 October 2012]. Available from: http://tinyurl.com/oo58stg

Network Rail. 2012e. *The 2014 Network Statement Final Version. Dated 31/10/12*. [Online] [Accessed 03 September 2013] Available from: http://tinyurl.com/ovwtjxv

Network Rail. 2013a. *A Better Railway for a Better Britain* , [Online] London, UK. Network Rail.  [Accessed 13 August 2013]. Available from: http://tinyurl.com/nj322c9

Network Rail. 2013b. *Network Rail Infrastructure Limited. Regulatory Financial Statements 2013. Year Ended March 2013*.[Online] [Accessed 06 September 2013] Available from: http://tinyurl.com/laeutdh

Network Rail. 2013c. *Periodic Review 2013 – Capacity Charge Conclusions & Draft Pricelists- April 2013*. [Online] [Accessed 06 September 2013] Available from: http://tinyurl.com/p2vosdb

Network Rail. 2013d. Rail Industry Seminar – Trade Offs. Analytical Framework and Toolkit [Powerpoint presentation] 23rd May 2013.Coventry, UK.

Network Rail. 2013e. RUS Documents.  [Online] [Accessed 03 September 2013] Available from: http://tinyurl.com/mqwqawz

Network Rail. 2013f. *Timetable Planning Rules. London North Eastern Version Issued 25 October 2013*.Milton Keynes, UK: Network Rail

Network Rail. 2013g. Trade Offs Summary Document[Online] [Accessed on 27 August 2013] Available from: http://tinyurl.com/omzezlh

Network Rail. 2014a. *London North Eastern Route – Sectional Appendix* Network Rail: York, UK [Accessed 08 August 2014]

Network Rail, 2014b. *Network Code – Part D* [Online] [Accessed 08 August 2014] Available from: http://tinyurl.com/m5h9l4o

Network Rail. 2014c. *Network Rail Infrastructure Limited. Regulatory Financial Statements 2014. Year Ended March 2014*.. [Online] [Accessed 24 October 2014] Available from: http://tinyurl.com/o9nu4eu

Network Rail. 2015. *Performance and Punctuality* [Online] [Accessed 09 January 2015] Available from: http://tinyurl.com/n7es5tx

Nichols Group. 2013. *HLOS Performance and Reliability Analysis and Targets*. Report for Network Rail and the Office of Rail Regulation. Final Report Dated 22/04/13. [Online] ORR: London, UK [Accessed 28/08/13] Available from: http://tinyurl.com/o7ghe8g

Nilson, J.E. 2002. Towards a welfare enhancing process to manage railway infrastructure access *Transportation Research Part A Volume **36** (2002)* pp 419-436.

Oliveira, E. and  Smith, B.M., 2000*. A Job-Shop Scheduling Model for the Single-Track Railway Scheduling Problem,* Research Report 2000.21, Leeds, UK: University of Leeds.

Olsson, N.O.E. and Haugland, H. Influencing factors on train punctuality—results from some Norwegian studies. *Transport policy* **11.4** (2004) pp387-397.

ORR. 2004a. Guide to the *Model Passenger Track Access Contract* [Online] London, UK: ORR [Accessed11 April 2013] Available from: http://tinyurl.com/ma84oek

ORR. 2004b Moderation of Competition : Final Conclusions. [Online] London, UK: ORR. [Accessed 11 April 2013] . Available from: http://tinyurl.com/nu96vle

ORR. 2009. *Promoting Safety and Value in Britain's Railways (Our plan for 2009-10, Year One of Our Strategy* [Online] London, UK: ORR. [Accessed 11April 2013]. Available from: http://tinyurl.com/qxoxber

ORR. 2010. *Review of Industry Approach to Capacity and Access Planning – Final Conclusions*. [Online] London, UK: ORR [Accessed 11 April2013] Available from: http://tinyurl.com/mfd7do8

ORR. 2012a. *Official Statistics obtained using ORR Data Portal.* [Online] London, UK : Office of Rail Regulation*.* [Accessed 13 December 2012] *Available from: www.dataportal.orr.gov.uk*

ORR. 2012b. *Periodic Review 2013 Consultation on Schedules 4 and 8 possessions and performance regimes* [Online] London, UK; ORR. [Accessed 27 August2013]. Available from: http://tinyurl.com/pvryedn

ORR. 2012c. *Annex C. Network Rail's Long Distance Sector Improvement Plan : Evidence Report following ORR's investigation.[*Online] London, UK; ORR. [Accessed 10 September 2013] Available from: http://tinyurl.com/ox7qggd

ORR. 2013a. *Official Statistics obtained using ORR Data Portal.* [Online] London, UK : Office of Rail Regulation*.* [Accessed 25th April 2013] *Available from: www.dataportal.orr.gov.uk*

ORR. 2013b. *Periodic Review 2013: Draft Determination of Network Rail's outputs and funding for 2014-19*. [Online] London, UK; ORR [Accessed 10 September 2013]. Available from: http://tinyurl.com/kp8gm92

ORR. 2013c. *Periodic Review 2013: Final Determination of Network Rail's outputs and funding for 2014-19*.[Online] London, UK: ORR. [Accessed 10 November 2013].Available from: http://tinyurl.com/pas2jlt

ORR. 2014.*The Network Code. [Online] [Accessed 28 March 2014] Available from:* http://tinyurl.com/qxooz28

ORR. 2015. *Track Access Guidance: Performance Regime. [Online] [Accessed 12 May 2015] Available from:* http://tinyurl.com/kh7g6ln

Parliament. 2008. *Reducing Rail Delays* [Online] [Accessed 15 August 2013] Available from: http://tinyurl.com/kjc6yft

Petersen E.R. 1974. Over the Road Transit Time for a Single-Track Railway*. Transportation Science.* Volume **8**(1) p65-74.

Preston, J. 1999. *Competition for Long Distance Passenger Rail Services; The Emerging Evidence*. [Discussion Paper No.2009-23 December 2009]. [no place] OECD/ITF pp1-23

Preston, J., Wall, G., Batley, R., Ibáñez, J. and Shires, J. 2009. Impact of delays on passenger train services. *Transportation Research Record: Journal of the Transportation Research Board*, ***2117***(1), pp14-23.

PTEG. 2012. Consultation Response. Network Rail Consultation on the Capacity Charge*.* [Online] [Accessed 17 October2012]. Available from: http://tinyurl.com/orfcdoj

QMS (Quantative Micro Software). 2010. *EViews 7 User's Guide II* [Online]. Irvine, USA: QMS Available from: http://tinyurl.com/lcqlkyy

Rail Freight Group. 2012. Periodic Review 2013 – Consultation on the Capacity Charge. Response from Rail Freight Group, August 2012. [Accessed on 17/10/12]. Available from: http://tinyurl.com/maeyc29

The Railways Infrastructure Access and Management Regulations 2005. SI 2005/3049 London: The Stationery Office [Online] [Accessed 03 September 2013] Available from: http://tinyurl.com/oad4uc9

The Railways Infrastructure Access and Management Regulations (Amended) 2009. SI 2009/222 London: The Stationery Office [Online] [Accessed 03 September 2013] Available from: http://tinyurl.com/kej4eo7

Railway-Technical.com. 2013 *Acronyms and Abbreviations for Railways* Railway Technical Pages. [Accessed on 27/08/13]. Available from www.railway-technical.com

Rouwendal, J. and Verhoef, E.T. 2006 *Basic Economic Principles of Road Pricing: From Theory to Applications Transport Policy* **13**(2). pp 106-114.

Schittenhelm, B and Landex, A. 2013 Development and Application of Danish Key Performance Indicators for Railway Timetables. *5th International Seminar on Railway Operations Modelling and Analysis, 13 to 15 May 2013*. *Copenhagen, Technical University of Denmark.* [no publisher] [no place] pp844-864

Sogin, S., Barkan, C. and Saat, M. 2011, Simulating the effects of higher speed passenger trains in single track freight networks [Online] *2011 Winter Simulation Conference. 11 to 14 December 2011, Phoenix, USA.* [Accessed 13/09/13] [no publisher] [no place] pp3684-3692 Available from: http://tinyurl.com/onwsrr5

Smithers, A. 2013. *National Rail Network Diagram*. [Online] Project Mapping [Accessed 19/05/15]. Available from: http://tinyurl.com/prd8nwe

Symonds Group Ltd. 2000. *Assessment of Capacity Charges – Final Report*. [Online] London, UK; Network Rail [Accessed 01/09/12]. Available from: http://tinyurl.com/lhb5ew2

Thomas, J. and McMahon, P. 2005 *Access Pricing in Rail – Principles and Structures* In: Vass.P. ed. Centre for the Study of Regulated Industries Conference '*Access Pricing , Investment and Efficient Use of Capacity in Network Industries', 8 December 2004, London, UK*. Bath; University of Bath School of Management. pp79-99.

Thompson, G., Hawkins, O., Dar, A. and Taylor, M. (eds). 2012. *Olympic Britain, Social and Economic Change since the 1908 and 1948 London Games*. [Online] London, UK; House of Commons Library. [Accessed 01 August 2013]. Available from: http://tinyurl.com/m4rbf7g

Trackmaps 2005 *Railway Track Diagrams, Eastern. 3rd ed. Quail Track Diagrams.* Bradford-on-Avon. UK: Trackmaps

Transport for London. 2012. PR13 – Consultation on the Capacity Charge [Online] [Accessed 17 October2012]. Available from: http://tinyurl.com/l3as7je

Transport Select Committee. 2013. *Rail 2020*, *Seventh Report*, *Volume 1* [Online] London, UK: Houses of Parliament. [Accessed 15 August 2013] Available from: http://tinyurl.com/nybrpc3

Turvey, R. 2000. Infrastructure Access Pricing and Lumpy Investments *Utilities Policy* **9**(2000) pp 207-218.

UIC (Union Internationale des Chemins de Fer), 2004. *Capacity UIC Code 406R 1st Edition September 2004,Paris, France.* International Union of Railways.

Vickrey, W.S. 1961. Congestion Theory and Transport Investment. *The American Economic Review*, **59**(2), Papers and Proceedings of the Eighty-first Annual Meeting of the American Economic Association (May, 1969), pp251-260

Vromans, M.J.C.M. 2005. *Reliability of Railway Systems* Ph.D Thesis, Erasmus University, Rotterdam.

Vromans, M.J.C.M., Dekker, R. and Kroon, L.G. 2006. Reliability and heterogeneity of railway services *European Journal of Operational Research* **172**(2006) pp647-655

Watson, R. 2008. *Train Planning in a Fragmented Railway – A British Perspective*. Ph.D Thesis. Loughborough University, UK.

Whelan, G., and Johnson, D. 2004. Modelling the impact of alternative fare structures on train overcrowding*. International journal of transport management, 2*(1), 51-58.

Wolmar, C. 1996. *The Great British Railway Disaster (The Independent on Sunday)*, Shepperton, UK: Ian Allan Publishing.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, Massachusetts, USA: The MIT Press,

Yuan J. and Hansen I. 2007. Optimising Capacity Utilisation at Stations by Estimating Knock-on Train Delays *Transportation Research Part B***41**(2) pp202-217

# List of Abbreviations

(Note : a substantial number of abbreviations are for analytical measures used in this thesis. Their definition can be found in Chapter Three).

| | |
|---|---|
| Adj. R-sq | Adjusted R-squared |
| ADT | ‘Average Distance Travelled’ (a measure used in this thesis) |
| AEL | ‘Average Entry Lateness’ (a measure used in this thesis) |
| AHET | Arrival HET. One of the Capacity Utilisation Measures tested in this Thesis. |
| Approx. | Approximation. Used as part of 2$^{nd}$ Order Approximation (a functional form used in this thesis). |
| ATV | ‘Average Train Time Variation’ (a measure used in this thesis) |
| Coeff. | Coefficient |
| CP4 | Control Period 4. The rail industry’s fourth control period which is from 2009 to 2014. |
| CP5 | Control Period 5. The rail industry’s fifth control period which is from 2014 to 2019. |
| CRRD | Congestion Related Reactionary Delay. |
| CTS | Constant Traffic Sections. A definition used by Arup (2013) to describe the philosophy behind the geographic sections in the recalibration of the Capacity Charge. |
| CUI | Capacity Utilisation Index (used in this thesis). |
| DSL | Down Slow Line |
| DfT | Department for Transport |
| Dn | Down (Direction of Travel. By Convention this is generally away from London). |
| EC | East Coast Trains |
| ECML | East Coast Main Line. The main line in Britain linking Yorkshire, the North East of England and Scotland with London. |

| | |
|---|---|
| EDIN | Edinburgh |
| ERTMS | European Rail Traffic Management System. The co-ordination of rail traffic management systems across Europe. |
| ETCS | European Train Control System. Forms part of ERTMS and features ‘in-cab’ signalling. |
| FE | Fixed effects. |
| FL | Fast line |
| FOC | Freight Operating Company |
| GC | Grand Central Trains |
| GLAS | Glasgow |
| GN | Great Northern. Trains operating as part of the Thameslink Great Northern Franchise. |
| Gthm | Grantham |
| GWML | Great Western Main Line. The main line in Britain linking the South West of the country with London. |
| HET | Methodology based on the Heterogeneity measures proposed by Vromans, Dekker and Kroon (2006) (used in this thesis). |
| HLOS | High Level Output Statement. The level of capability the Government wishes to see. |
| HS2 | The proposed (at the time of writing) new high speed line linking London with the north of Britain via Birmingham. |
| HT | Hull Trains |
| I | Intensity (of trains – a Capacity Utilisation Measure used in this Thesis) |
| ICEC | Intercity East Coast. |
| IEP | Intercity Express Programme. The next generation of Intercity trains. |
| LDS | Leeds |
| Log. | Logarithmic |
| MP | Member of Parliament |
| MSX | Monday / Saturday Excepted (Timetable Designation) i.e. the train path is Tuesday to Friday. |

| | |
|---|---|
| NWC | Newcastle |
| NMF | Network Modelling Framework. A detailed strategic and forecasting appraisal model. |
| NR | Network Rail. |
| OCUI | Original CUI. One of the capacity utilisation measures tested in this thesis. |
| OHET | Original HET. One of the capacity utilisation measures tested in this thesis. |
| ORR | Office of Rail Regulation. The regulating body for the rail industry in Britain. |
| PPM | Public Performance Measure. A measure of performance intended to show the ‘success’ of the rail industry in delivering an acceptable level of performance to customers. |
| PRAISE | Privatised Rail Services model. The software encompasses a demand element, a cost element and an evaluation element. |
| RD1TM | (Congestion Related Reactionary Delay) + 1 / Train Miles – The Standard Dependent Variable used in the regression analysis carried out for this thesis. |
| RE | Random effects. |
| RFOA | Rail Freight Operator’s Association. |
| RPI | Retail Price Index. |
| RUS | Route Utilisation Strategy. |
| SBCAP | Capacity Utilisation in the Section Before (a measure used in this thesis). |
| SBCUI | CUI in the Section Before (a measure used in this thesis). |
| SC | Service Code. |
| SFCAP | Capacity Utilisation in the Section Following (a measure used in this thesis). |
| SFCUI | CUI in the Section Following (a measure used in this thesis). |
| SL | Slow line |
| SoFA | Statement of Funds Available. The amount of public funding to be made available to facilitate delivery of the HLOS. |

| | |
|---|---|
| SRT | Sectional Running Time. |
| STAB | ‘Stability’ (a measure used in this thesis) |
| SUND | Sunderland |
| SX | Saturday Excepted (Timetable designation) the train path is Monday to Friday. |
| TBCAP | Capacity Utilisation in the Time Period Before ( a measure used in this thesis). |
| TBCUI | Calculated CUI in the Time Period Before ( a measure used in this thesis). |
| TOCs | (Passenger) Train Operating Companies |
| TOPS | Train Operating System |
| TRUST | Train Running System on TOPS (a performance monitoring system in use on Britain’s rail network). |
| TTC | ‘Timetable Complexity’ (a measure used in this thesis) |
| TThFO | Tuesday Thursday Friday Only (Timetable Designation) i.e. the train paths are planned to operate as indicated. |
| t-stat | t-statistic. |
| Up | Up (Direction of Travel. By convention this is generally towards London). |
| VHETB | Vulnerable HET (Gap Before). One of the capacity utilisation measures tested in this thesis. |
| VHETF | Vulnerable HET (Gap Following). One of the capacity utilisation measures tested in this thesis. |
| WCML | West Coast Main Line. The main line in Britain linking the North West of England and Glasgow with London. |
| WO | Wednesday Only (Timetable Designation) i.e. the train paths are planned to operate as indicated. |
| XCUI | Junction CUI. One of the capacity utilisation measures tested in this thesis. |
| XHET | Junction HET. One of the capacity utilisation measures tested in this thesis. |

# Appendix A
# Freight Paths Included in the Analysis

This appendix shows the weekly freight paths that were included in the analysis based on the amount of times they actually operated. Paths which operated less than 5% of the days in the weekday December 2008 to May 2009 Timetable were excluded. As previously discussed paths which were only planned for a single day of the week (e.g. Wednesday Only or WO) were excluded prior to this analysis and are not shown here.

**Table A.1** Newark 'Down' Freight Paths Included in the Analysis

| Train | TT | Path | % Run | Included |
|---|---|---|---|---|
| 4E58 | MSX | Felixstowe to Leeds | 71.8 | YES |
| 4E78 | MSX | Felixstowe North to Selby | 72.7 | YES |
| 6H92 | SX | Peterboro W Yd to Goole Glass Wks | 30.9 | YES |
| 6E45 | MSX | Felixstowe Sth to Wakefield Europt | 2.7 | no |
| 4E28 | MSX | Tilbury Cont. to Wakefield Europt | 48.2 | YES |
| 4E62 | SX | Ipswich to Leeds FLiner Terminal | 1.8 | no |
| 6E84 | SX | Middleton to Barnby / Monk Bretton | 68.2 | YES |
| 6E82 | SX | Rectory to Lindsey | 93.6 | YES |
| 4E24 | SX | Grain Thamesport to Leeds FLiner | 92.7 | YES |
| 4E33 | SX | Felixstowe to Doncaster | 90.0 | YES |
| 4E19 | SX | Mountfield to West Burton | 41.8 | YES |
| 4E32 | SX | Dollands Moor to Scunthorpe | 60.9 | YES |
| 4D56 | SX | Biggleswade to Heck | 29.1 | YES |
| 4E55 | SX | Felixstowe to Doncaster | 90.9 | YES |
| 6E83 | SX | Ketton Ward to Lindsey | 1.8 | no |
| 4E50 | SX | Felixstowe to Leeds FLiner Terminal | 91.8 | YES |

**Table A.2** Freight Paths Crossing the ECML at Newark Flat Crossing Included in the Analysis

| Train | TT | Path | % Run | Included |
|---|---|---|---|---|
| 6E46 | MSX | Kingsbury to Lindsey | 46.4 | YES |
| 6M57 | SX | Lindsey to Kingsbury | 68.2 | YES |
| 4M82 | SX | West Burton to Hotchley Hill | 0.0 | no |
| 6M00 | SX | Humber to Kingsbury | 57.3 | YES |
| 6E54 | SX | Kingsbury to Humber | 86.4 | YES |
| 6M88 | SX | Immingham to Ketton | 0.0 | no |
| 6A59 | SX | Hatfield Colliery to Ratcliffe | 3.6 | no |
| 6E21 | MSX | Mountsorrell to Ratcliffe | 0.9 | no |
| 6E98 | WThFO | Daw Mill to Drax | 0.0 | no |
| 6E41 | SX | Westleigh Murco to Lindsey | 79.1 | YES |
| 6E59 | SX | Kingsbury to Lindsey | 76.4 | YES |
| 6M24 | SX | Lindsey to Kingsbury | 66.4 | YES |
| 6E38 | SX | Colnbrook to Lindsey Oil | 76.4 | YES |
| 6E55 | MWFO | Theale Murco to Lindsey Oil | 42.7 | YES |

**Table A.3** Newark 'Up' Freight Paths Included in the Analysis

| Train | TT | Path | % Run | Included |
|---|---|---|---|---|
| 4O20 | MSX | West Burton to Mountfield | 0.0 | no |
| 4L45 | SX | Wakefield to Felixstowe | 88.2 | YES |
| 4L85 | SX | Leeds to Felixstowe | 94.5 | YES |
| 4L78 | SX | Selby to Felixstowe | 93.6 | YES |
| 6L55 | SX | Wakefield to Felixstowe | 2.5 | no |
| 4L28 | SX | Wakefield to Tilbury | 65.5 | YES |
| 6H93 | SX | Goole Glass to Peterboro West Yd | 30.9 | YES |
| 4L79 | SX | Wilton to Felixstowe | 87.3 | YES |
| 6O19 | SX | Scunthorpe to Dollands Moor | 59.1 | YES |
| 6L84 | SX | Doncaster to Whitemoor Yard | 81.8 | YES |
| 4L64 | SX | Leeds FLiner Terminal to Tilbury | 72.7 | YES |
| 6D28 | SX | Barnby Dunn to Peterborough | 66.4 | YES |

**Table A.4** Welwyn 'Down' Freight Paths Included in the Analysis

| Train | TT | Path | % Run | Included |
|---|---|---|---|---|
| 6M57 | TThFO | Hitchin to Peak Forest | 0.0 | no |
| 6M67 | WFO | Broxbourne to Mount Sorrell | 10.9 | YES |
| 4E19 | SX | Mountfield to West Burton | 41.8 | YES |
| 4E32 | SX | Dollands Moor to Scunthorpe | 60.0 | YES |
| 4E85 | SX | Tilbury to Belmont | 0.0 | no |
| 6E52 | TThO | Cardiff Tidal to Hitchin Up Yard | 3.6 | no |
| 4D56 | SX | Bigglesw Plasmor to Heck Plasmor | 23.6 | YES |
| 4E24 | SX | Grain Thamesport to Leeds | 91.8 | YES |
| 4E25 | SX | Bow Depot to Heck Plasmor | 59.1 | YES |

**Table A.5** Welwyn ‘Up’ Freight Paths Included in the Analysis

| Train | TT | Path | % Run | Included |
|---|---|---|---|---|
| 6M57 | TThFO | Hitchin Up Yd to Peak Forest Sdgs | 0.0 | no |
| 6L69 | SX | Peterboro West Yd to Bow Depot | 57.3 | YES |
| 0M65 | SX | Peterboro Maint Shed to Wembley | 0.0 | no |
| 4O20 | SX | West Burton to Mountfield Sdgs | 0.0 | no |
| 4L45 | SX | Wakefield Europort to Felixstowe | 89.1 | YES |
| 6V52 | TThO | Hitchin to Acton Yard | 3.6 | no |
| 4L28 | SX | Wakefield Europort to Tilbury | 64.5 | YES |

# Appendix B

## Sample of the Data Set

**Table B.1** Sample of the Sectional Data Set for Stevenage to Woolmer (Up Fast): Capacity Utilisation Explanatory Variables.

| **Data** | **0600-0700** | **0700-0800** | **0800-0900** | **0900-1000** | **1000-1100** | **1100-1200** |
|---|---|---|---|---|---|---|
| CRRD | 12 | 49 | 95 | 57 | 73 | 6 |
| Train Miles | 25.97 | 33.39 | 29.68 | 29.68 | 29.68 | 29.68 |
| RD1 | 13.0 | 50.0 | 96.0 | 58.0 | 74.0 | 7.0 |
| **RD1TM** | 0.50 | 1.50 | 3.23 | 1.95 | 2.49 | 0.24 |
| **I** | 35.0% | 45.0% | 40.0% | 40.0% | 40.0% | 40.0% |
| **OCUI** | 49.2% | 49.2% | 42.5% | 45.0% | 46.7% | 50.0% |
| **XCUI** | 49.2% | 66.7% | 76.7% | 61.7% | 63.3% | 53.3% |
| **OHET** | 51.8% | 48.5% | 51.1% | 48.2% | 57.6% | 54.2% |
| **AHET** | 47.1% | 47.2% | 50.0% | 45.9% | 53.5% | 51.3% |
| **XHET** | 55.7% | 55.8% | 66.5% | 54.1% | 76.7% | 59.6% |
| **VHETB** | 51.8% | 48.5% | 48.9% | 48.2% | 61.1% | 54.2% |
| **VHETF** | 51.8% | 48.5% | 51.1% | 46.8% | 52.7% | 54.2% |

**Table B.2** Sample of the Sectional Data Set for Stevenage to Woolmer (Up Fast): 'Other' Explanatory Variables.

| **Data** | **0600-0700** | **0700-0800** | **0800-0900** | **0900-1000** | **1000-1100** | **1100-1200** |
|---|---|---|---|---|---|---|
| CRRD | 12 | 49 | 95 | 57 | 73 | 6 |
| Train Miles | 25.97 | 33.39 | 29.68 | 29.68 | 29.68 | 29.68 |
| RD1 | 13.0 | 50.0 | 96.0 | 58.0 | 74.0 | 7.0 |
| **RD1TM** | 0.50 | 1.50 | 3.23 | 1.95 | 2.49 | 0.24 |
| **STAB (minutes)** | 0.000 | 0.167 | 0.188 | 0.000 | 0.563 | 0.250 |
| **TTC (number of SCs)** | 4 | 5 | 5 | 5 | 6 | 4 |
| **ATV (minutes)** | 1.821 | 1.194 | 1.094 | 1.188 | 1.469 | 1.438 |
| **AEL (minutes)** | 0.694 | 1.626 | 3.413 | 3.128 | 3.448 | 1.887 |
| **ADT (miles)** | 40.49 | 95.99 | 155.39 | 185.53 | 181.56 | 150.92 |
| **TBOCUI** [48] | 0.0% | 49.2% | 49.2% | 42.5% | 45.0% | 46.7% |
| **SBOCUI** | 42.5% | 48.3% | 42.5% | 45.0% | 45.0% | 45.8% |
| **SFOCUI** | 44.2% | 83.3% | 86.7% | 73.3% | 66.7% | 56.7% |

[48] For brevity only data for OCUI given.

**Table B.3** Sample of the Area Data Set for Welwyn Up Fast: Capacity Utilisation Explanatory Variables.

| **Data** | **0600-0700** | **0700-0800** | **0800-0900** | **0900-1000** | **1000-1100** | **1100-1200** |
|---|---|---|---|---|---|---|
| CRRD | 102.0 | 414.0 | 716.0 | 556.0 | 765.0 | 248.4 |
| Train Miles | 242.54 | 398.11 | 492.11 | 419.97 | 379.53 | 328.68 |
| RD1 | 103.0 | 415.0 | 717.0 | 557.0 | 766.0 | 249.4 |
| **RD1TM** | 0.42 | 1.04 | 1.45 | 1.33 | 2.02 | 0.76 |
| **LCUI** | 44.2% | 83.3% | 86.7% | 73.3% | 66.7% | 56.7% |
| **LHET** | 53.6% | 83.5% | 87.6% | 79.0% | 83.8% | 71.4% |
| **EHET** | 49.2% | 79.3% | 87.6% | 84.1% | 76.5% | 68.5% |

# Appendix C
# Residual Sums of the Squares

This appendix gives the residual sums of the squares results for the three options excluded from the main text of the thesis in the interests of brevity. These are two-way fixed effects; one-way random effects and two-way random Effects. It can be seen that in the case of the sectional capacity variables, the non-linear (logarithmic) functional forms are preferred to the linear ones (i.e. they have smaller residual sums of squares). In contrast the linear functional forms are preferred for the area capacity variables.

A number of specifications could not be calculated using the Second Order Approximation (logarithmic) functional form due to perfect colinearity. These are identified by a *.

**Table C.1** Comparison of the Residual Sums of Squares for the Five Functional Forms (Two-Way / Fixed Effects).

| Capacity Utilisation Variable | Linear (linear) | Quadratic (linear) | 2$^{nd}$ Order Approx. (linear) | Exponential (logarithmic) | 2$^{nd}$ Order Approx. (log.) |
|---|---|---|---|---|---|
| Intensity | 575.95 | 573.72 | 569.56 | 305.92 | * |
| OCUI | 572.49 | 571.42 | 571.05 | 302.75 | 301.51 |
| XCUI | 569.13 | 566.79 | 565.20 | 301.92 | 298.85 |
| OHET | 546.23 | 550.37 | 545.81 | 278.49 | 281.46 |
| AHET | 562.65 | 567.29 | 560.11 | 290.80 | 289.39 |
| XHET | 535.05 | 540.50 | 533.48 | 274.42 | 270.04 |
| VHETB | 544.95 | 548.82 | 544.62 | 280.22 | 281.64 |
| VHETF | 551.58 | 555.30 | 551.03 | 282.12 | 283.35 |
| LCUI | 8.92 | 9.25 | 8.24 | 10.46 | * |
| LHET | 9.21 | 9.19 | 9.18 | 10.26 | * |
| EHET | 7.99 | 8.13 | 7.79 | 9.67 | * |

**Table C.2** Comparison of the Residual Sums of Squares for the Five Functional Forms (One-Way / Random Effects).

| **Capacity Utilisation Variable** | **Linear (linear)** | **Quadratic (linear)** | **2$^{nd}$ Order Approx. (linear)** | **Exponential (logarithmic)** | **2$^{nd}$ Order Approx. (log.)** |
|---|---|---|---|---|---|
| Intensity | 638.88 | 636.19 | 633.48 | 344.08 | * |
| OCUI | 635.49 | 635.26 | 634.22 | 343.24 | 344.31 |
| XCUI | 633.29 | 631.40 | 629.37 | 343.06 | 341.00 |
| OHET | 603.35 | 607.27 | 603.83 | 305.44 | 309.03 |
| AHET | 619.75 | 625.98 | 619.61 | 322.21 | 319.49 |
| XHET | 594.53 | 599.07 | 592.66 | 301.21 | 294.38 |
| VHETB | 601.40 | 605.04 | 601.59 | 307.76 | 309.59 |
| VHETF | 610.60 | 614.87 | 610.84 | 311.45 | 312.34 |
| LCUI | 12.39 | 12.89 | 11.49 | 15.70 | * |
| LHET | 12.71 | 12.82 | 12.99 | 15.84 | * |
| EHET | 10.50 | 10.68 | 11.77 | 14.61 | * |

**Table C.3** Comparison of the Residual Sums of Squares for the Five Functional Forms (Two-Way / Random Effects).

| **Capacity Utilisation Variable** | **Linear (linear)** | **Quadratic (linear)** | **2nd Order Approx. (linear)** | **Exponential (logarithmic)** | **2nd Order Approx. (log.)** |
|---|---|---|---|---|---|
| Intensity | 638.68 | 634.95 | 631.59 | 343.78 | * |
| OCUI | 633.45 | 631.93 | 632.94 | 341.79 | 343.70 |
| XCUI | 630.45 | 627.78 | 625.33 | 339.14 | 340.51 |
| OHET | 603.47 | 607.41 | 602.86 | 305.72 | 309.31 |
| AHET | 619.96 | 626.12 | 619.83 | 322.30 | 319.66 |
| XHET | 594.32 | 599.21 | 591.19 | 301.39 | 294.64 |
| VHETB | 601.54 | 605.21 | 601.21 | 307.98 | 309.82 |
| VHETF | 609.51 | 613.51 | 608.25 | 311.58 | 312.52 |
| LCUI | 12.05 | 12.52 | 10.93 | 14.49 | * |
| LHET | 12.72 | 12.67 | 12.99 | 14.62 | * |
| EHET | 10.42 | 10.60 | 11.65 | 14.42 | * |

# Appendix D Detailed Regression Results

**Table D.1** Regression Results for Intensity (I) (Exponential One- Way FE)

| Area | Geographic Section | Coeff. | t-statistic (White) | Adjust. R-sq. |
|---|---|---|---|---|
| Intensity | β | 0.02993 | 5.054 (4.314) | 0.432 |
| Newark Dn | Grantham - Newark | 0.33374 | -3.44536 | |
| | Newark - Retford | 0.13896 | -2.60488 | |
| | Retford - Loversall | 0.10631 | -3.40152 | |
| Newark Up | Loversall – Retford | 0.34475 | -2.47977 | |
| | Retford - Newark | 0.20065 | -1.51476 | |
| | Newark - Grantham | 0.27854 | -0.53823 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.07356 | -0.42411 | |
| | Potters Bar to Welwyn (SL) | 0.15184 | -2.34165 | |
| | Welwyn Viaduct | 0.23820 | -0.93096 | |
| | Woolmer – Stevenage (FL) | 0.25390 | -0.81302 | |
| | Woolmer – Stevenage (SL) | 1.61429 | 2.37016 | |
| | Stevenage - Hitchin (FL) | 0.45557 | 0.92529 | |
| | Stevenage - Hitchin (SL) | 0.45104 | 0.89591 | |
| | Hitchin - Sandy (FL) | 0.03903 | -6.38505 | |
| | Hitchin - Sandy (SL) | 0.09560 | -3.47396 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.31964 | -0.12850 | |
| | Sandy – Hitchin (SL) | 1.02281 | 3.12243 | |
| | Hitchin – Stevenage (FL) | 0.33123 | -0.02245 | |
| | Hitchin – Stevenage (SL) | 0.14485 | -2.48177 | |
| | Stevenage – Woolmer (FL) | 0.35990 | 0.22435 | |
| | Stevenage – Woolmer (SL) | 0.68581 | 2.14156 | |
| | Welwyn Viaduct | 0.32803 | -0.04727 | |
| | Welwyn – Potters Bar (FL) | 0.04827 | -5.62612 | |
| | Welwyn – Potters Bar (SL) | 0.26002 | -0.74296 | |

**Table D.2** Regression Results for OCUI (Exponential One- Way FE)

| **Area** | **Geographic Section** | **Coeff.** | **t-statistic (White)** | **Adjust. R-sq.** |
|---|---|---|---|---|
| OCUI | β | 0.02376 | 5.124 (4.378) | 0.433 |
| Newark Dn | Grantham - Newark | 0.30674 | -3.60779 | |
| | Newark - Retford | 0.13943 | -2.33932 | |
| | Retford - Loversall | 0.10740 | -3.11001 | |
| Newark Up | Loversall – Retford | 0.14244 | -2.28438 | |
| | Retford - Newark | 0.18818 | -1.45601 | |
| | Newark - Grantham | 0.24673 | -0.64885 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.09139 | -3.60826 | |
| | Potters Bar to Welwyn (SL) | 0.13915 | -2.35249 | |
| | Welwyn Viaduct | 0.30222 | -0.04330 | |
| | Woolmer – Stevenage (FL) | 0.30036 | -0.06238 | |
| | Woolmer – Stevenage (SL) | 0.78093 | 2.78419 | |
| | Stevenage - Hitchin (FL) | 0.47652 | 1.31238 | |
| | Stevenage - Hitchin (SL) | 0.49761 | 1.42589 | |
| | Hitchin - Sandy (FL) | 0.04594 | -5.56855 | |
| | Hitchin - Sandy (SL) | 0.09910 | -3.08474 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.35322 | 0.41696 | |
| | Sandy – Hitchin (SL) | 1.06106 | 3.38467 | |
| | Hitchin – Stevenage (FL) | 0.37057 | 0.56044 | |
| | Hitchin – Stevenage (SL) | 0.17431 | -1.67491 | |
| | Stevenage – Woolmer (FL) | 0.38306 | 0.66041 | |
| | Stevenage – Woolmer (SL) | 0.74588 | 2.64519 | |
| | Welwyn Viaduct | 0.44940 | 1.11839 | |
| | Welwyn – Potters Bar (FL) | 0.05693 | -5.00720 | |
| | Welwyn – Potters Bar (SL) | 0.22225 | -0.95820 | |

**Table D.3** Regression Results for XCUI (Exponential One-Way FE)

| Area | Geographic Section | Coeff. | t-statistic (White) | Adjust. R-sq. |
|---|---|---|---|---|
| XCUI | β | 0.02452 | 5.276 (4.597) | 0.435 |
| Newark Dn | Grantham - Newark | 0.29556 | -3.72123 | |
| | Newark - Retford | 0.10264 | -3.15274 | |
| | Retford - Loversall | 0.10409 | -3.09893 | |
| Newark Up | Loversall – Retford | 0.13754 | -2.28278 | |
| | Retford - Newark | 0.14673 | -2.07733 | |
| | Newark - Grantham | 0.23479 | -0.68722 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.08079 | -3.86621 | |
| | Potters Bar to Welwyn (SL) | 0.10778 | -2.96766 | |
| | Welwyn Viaduct | 0.28783 | -0.07746 | |
| | Woolmer – Stevenage (FL) | 0.29095 | -0.04676 | |
| | Woolmer – Stevenage (SL) | 0.75337 | 2.79349 | |
| | Stevenage - Hitchin (FL) | 0.24846 | 0.64991 | |
| | Stevenage – Hitchin (SL) | 0.38745 | 1.45309 | |
| | Hitchin - Sandy (FL) | 0.04493 | -5.55006 | |
| | Hitchin - Sandy (SL) | 0.06000 | -4.69977 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.24846 | -0.51769 | |
| | Sandy – Hitchin (SL) | 0.38745 | 0.80257 | |
| | Hitchin – Stevenage (FL) | 0.35907 | 0.57823 | |
| | Hitchin – Stevenage (SL) | 0.16894 | -1.66092 | |
| | Stevenage – Woolmer (FL) | 0.27596 | -0.20399 | |
| | Stevenage – Woolmer (SL) | 0.53276 | 1.74618 | |
| | Welwyn Viaduct | 0.42853 | 1.09000 | |
| | Welwyn – Potters Bar (FL) | 0.05465 | -5.02875 | |
| | Welwyn – Potters Bar (SL) | 0.21340 | -0.97068 | |

**Table D.4** Regression Results for OHET (Exponential One- Way FE)

| **Area** | **Geographic Section** | **Coeff Value** | **t-statistic (White)** | **Adjust. R-sq.** |
|---|---|---|---|---|
| OHET | β | 0.03830 | 8.714 (7.931) | 0.498 |
| Newark Dn | Grantham - Newark | 0.13516 | -6.28750 | |
| | Newark - Retford | 0.06338 | -2.39235 | |
| | Retford - Loversall | 0.04397 | -3.55342 | |
| Newark Up | Loversall – Retford | 0.06213 | -2.46090 | |
| | Retford - Newark | 0.10204 | -0.88732 | |
| | Newark - Grantham | 0.12124 | -0.34417 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.03379 | -4.37845 | |
| | Potters Bar to Welwyn (SL) | 0.09653 | -2.97873 | |
| | Welwyn Viaduct | 0.11199 | -1.03444 | |
| | Woolmer – Stevenage (FL) | 0.18013 | -0.59541 | |
| | Woolmer – Stevenage (SL) | 0.01910 | 1.42800 | |
| | Stevenage - Hitchin (FL) | 0.20583 | 0.90858 | |
| | Stevenage - Hitchin (SL) | 0.15962 | -0.52785 | |
| | Hitchin - Sandy (FL) | 0.16798 | -6.17020 | |
| | Hitchin - Sandy (SL) | 0.11879 | -1.70919 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.01945 | 1.38087 | |
| | Sandy – Hitchin (SL) | 0.21341 | 5.17050 | |
| | Hitchin – Stevenage (FL) | 0.07449 | 0.52580 | |
| | Hitchin – Stevenage (SL) | 0.05257 | -2.76900 | |
| | Stevenage – Woolmer (FL) | 0.27318 | 0.68762 | |
| | Stevenage – Woolmer (SL) | 0.08438 | 2.22647 | |
| | Welwyn Viaduct | 0.05632 | -0.39124 | |
| | Welwyn – Potters Bar (FL) | 0.11422 | -6.08280 | |
| | Welwyn – Potters Bar (SL) | 0.82880 | -1.48551 | |

**Table D.5** Regression Results for AHET (Exponential One-Way FE)

| Area | Geographic Section | Coeff. Value | t-statistic (White) | Adjust. R-sq. |
|---|---|---|---|---|
| AHET | β | 0.03677 | 7.308 (6.344) | 0.470 |
| Newark Dn | Grantham - Newark | 0.21074 | -5.01089 | |
| | Newark - Retford | 0.07984 | -2.99250 | |
| | Retford - Loversall | 0.05762 | -3.99687 | |
| Newark Up | Loversall – Retford | 0.08711 | -2.72329 | |
| | Retford - Newark | 0.12759 | -1.54726 | |
| | Newark - Grantham | 0.17897 | -0.50370 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.03944 | -5.06223 | |
| | Potters Bar to Welwyn (SL) | 0.12035 | -3.10979 | |
| | Welwyn Viaduct | 0.13612 | -1.61264 | |
| | Woolmer – Stevenage (FL) | 0.22090 | -1.34015 | |
| | Woolmer – Stevenage (SL) | 0.02265 | 0.35771 | |
| | Stevenage - Hitchin (FL) | 0.27081 | 0.14370 | |
| | Stevenage - Hitchin (SL) | 0.18479 | -1.32961 | |
| | Hitchin - Sandy (FL) | 0.20039 | -6.87677 | |
| | Hitchin - Sandy (SL) | 0.13492 | -2.81988 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.02401 | 0.76585 | |
| | Sandy – Hitchin (SL) | 0.23810 | 4.11877 | |
| | Hitchin – Stevenage (FL) | 0.07919 | -0.40443 | |
| | Hitchin – Stevenage (SL) | 0.07617 | -3.39349 | |
| | Stevenage – Woolmer (FL) | 0.49695 | -0.15496 | |
| | Stevenage – Woolmer (SL) | 0.14636 | 2.64450 | |
| | Welwyn Viaduct | 0.06920 | -1.23650 | |
| | Welwyn – Potters Bar (FL) | 0.13459 | -6.48863 | |
| | Welwyn – Potters Bar (SL) | 0.88995 | -1.12169 | |

**Table D.6** Regression Results for XHET (Exponential One-Way FE)

| Area | Geographic Section | Coeff. Value | t-statistic (White) | Adjust. R-sq. |
|---|---|---|---|---|
| XHET | β | 0.03939 | 8.971 (8.601) | 0.503 |
| Newark Dn | Grantham - Newark | 0.12778 | -6.48372 | |
| | Newark - Retford | 0.04373 | -3.40960 | |
| | Retford - Loversall | 0.04169 | -3.56209 | |
| Newark Up | Loversall – Retford | 0.05882 | -2.46867 | |
| | Retford - Newark | 0.08275 | -1.38266 | |
| | Newark - Grantham | 0.11273 | -0.39870 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.03042 | -4.55083 | |
| | Potters Bar to Welwyn (SL) | 0.08949 | -3.64225 | |
| | Welwyn Viaduct | 0.10587 | -1.09974 | |
| | Woolmer – Stevenage (FL) | 0.13611 | -0.59868 | |
| | Woolmer – Stevenage (SL) | 0.01819 | 1.39592 | |
| | Stevenage - Hitchin (FL) | 0.13121 | 0.19963 | |
| | Stevenage - Hitchin (SL) | 0.15155 | -0.56592 | |
| | Hitchin - Sandy (FL) | 0.11728 | -6.17939 | |
| | Hitchin - Sandy (SL) | 0.10966 | -5.09774 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.01819 | 0.19963 | |
| | Sandy – Hitchin (SL) | 0.19924 | -0.56592 | |
| | Hitchin – Stevenage (FL) | 0.02556 | 0.54226 | |
| | Hitchin – Stevenage (SL) | 0.04008 | -2.79496 | |
| | Stevenage – Woolmer (FL) | 0.17005 | -0.27235 | |
| | Stevenage – Woolmer (SL) | 0.07918 | 0.89457 | |
| | Welwyn Viaduct | 0.05305 | -0.46544 | |
| | Welwyn – Potters Bar (FL) | 0.10677 | -6.14740 | |
| | Welwyn – Potters Bar (SL) | 0.26160 | -1.51651 | |

**Table D.7** Regression Results for VHETB (Exponential One-Way FE)

| Area | Geographic Section | Coeff. Value | t-statistic (White) | Adjust. R-sq. |
|---|---|---|---|---|
| VHETB | | 0.03531 | 8.491 (7.574) | 0.493 |
| Newark Dn | Grantham - Newark | 0.14390 | -6.09665 | |
| | Newark - Retford | 0.06591 | -2.45686 | |
| | Retford - Loversall | 0.04961 | -3.35068 | |
| Newark Up | Loversall – Retford | 0.06673 | -2.4218 | |
| | Retford - Newark | 0.11712 | -0.64537 | |
| | Newark - Grantham | 0.14557 | 0.03623 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.03790 | -4.19982 | |
| | Potters Bar to Welwyn (SL) | 0.11322 | -2.66128 | |
| | Welwyn Viaduct | 0.12602 | -0.73847 | |
| | Woolmer – Stevenage (FL) | 0.20379 | -0.41832 | |
| | Woolmer – Stevenage (SL) | 0.02176 | 1.82263 | |
| | Stevenage - Hitchin (FL) | 0.23192 | 1.09675 | |
| | Stevenage - Hitchin (SL) | 0.18086 | -0.17585 | |
| | Hitchin - Sandy (FL) | 0.19309 | -5.91449 | |
| | Hitchin - Sandy (SL) | 0.15120 | -1.77854 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.02359 | 1.47328 | |
| | Sandy – Hitchin (SL) | 0.25756 | 5.04021 | |
| | Hitchin – Stevenage (FL) | 0.07720 | 0.71823 | |
| | Hitchin – Stevenage (SL) | 0.06179 | -2.46014 | |
| | Stevenage – Woolmer (FL) | 0.32007 | 0.92442 | |
| | Stevenage – Woolmer (SL) | 0.10051 | 2.52000 | |
| | Welwyn Viaduct | 0.06594 | 0.15144 | |
| | Welwyn – Potters Bar (FL) | 0.13606 | -5.67765 | |
| | Welwyn – Potters Bar (SL) | 0.84681 | -1.12985 | |

**Table D.8** Regression Results for VHETF (Exponential One-Way FE)

| Area | Geographic Section | Coeff. Value | t-statistic (White) | Adjust. R-sq. |
|---|---|---|---|---|
| VHETF | | 0.03487 | 8.190 (7.392) | 0.487 |
| Newark Dn | Grantham - Newark | 0.15841 | -5.82181 | |
| | Newark - Retford | 0.07342 | -2.40435 | |
| | Retford - Loversall | 0.04882 | -3.68800 | |
| Newark Up | Loversall – Retford | 0.07403 | -2.38325 | |
| | Retford - Newark | 0.11785 | -0.92427 | |
| | Newark - Grantham | 0.14176 | -0.34796 | |
| Welwyn Dn | Potters Bar to Welwyn (FL) | 0.04026 | -4.28313 | |
| | Potters Bar to Welwyn (SL) | 0.11948 | -2.80075 | |
| | Welwyn Viaduct | 0.12563 | -0.85923 | |
| | Woolmer – Stevenage (FL) | 0.20886 | -0.72665 | |
| | Woolmer – Stevenage (SL) | 0.02123 | 1.56758 | |
| | Stevenage - Hitchin (FL) | 0.25334 | 0.86542 | |
| | Stevenage - Hitchin (SL) | 0.19616 | -0.33097 | |
| | Hitchin - Sandy (FL) | 0.20212 | -6.27992 | |
| | Hitchin - Sandy (SL) | 0.16147 | -2.04387 | |
| Welwyn Up | Sandy – Hitchin (FL) | 0.02464 | 1.44002 | |
| | Sandy – Hitchin (SL) | 0.26269 | 4.85513 | |
| | Hitchin – Stevenage (FL) | 0.07755 | 0.66786 | |
| | Hitchin – Stevenage (SL) | 0.06465 | -2.65313 | |
| | Stevenage – Woolmer (FL) | 0.33187 | 0.76243 | |
| | Stevenage – Woolmer (SL) | 0.10308 | 2.31696 | |
| | Welwyn Viaduct | 0.06790 | 0.058017 | |
| | Welwyn – Potters Bar (FL) | 0.14242 | -5.79530 | |
| | Welwyn – Potters Bar (SL) | 0.87728 | -1.34206 | |

**Table D9** Regression Results for LCUI (Linear One-Way FE)

| Area | Coeff. Value | t-statistic | Adjusted R-squared |
|---|---|---|---|
| LCUI | 0.01295 | 2.98667 | 0.174 |
| Newark Down | 0.15448 | 0.69918 | |
| Newark Up | 0.26600 | 0.85442 | |
| Welwyn Down | 0.09625 | -0.38577 | |
| Welwyn Up | 0.32000 | 1.12284 | |

**Table D.10** Regression Results for LHET (Linear One-Way FE)

| Area | Coeff. Value | t-statistic | Adjusted R-squared |
|---|---|---|---|
| LHET | 0.01380 | 2.78655 | 0.160 |
| Newark Down | 0.06925 | 0.26309 | |
| Newark Up | 0.24965 | 1.36951 | |
| Welwyn Down | -0.03410 | -0.63005 | |
| Welwyn Up | 0.11420 | 0.25567 | |

**Table D.11** Regression Results for EHET (Linear One-Way FE)

| Area | Coeff. Value | t-statistic | Adjusted R-squared |
|---|---|---|---|
| EHET | 0.02026 | 4.64487 | 0.304 |
| Newark Down | -0.58321 | -1.94330 | |
| Newark Up | -0.34115 | 1.99682 | |
| Welwyn Down | -0.57146 | 0.09466 | |
| Welwyn Up | -0.27812 | 2.53640 | |