

**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Civil Engineering and the Environment

School of Electronics and Computer Science



**Rail Journey Recovery Following an Incident**

by

**John Armstrong**

Thesis for the degree of Doctor of Engineering

September 2004

## **ABSTRACT**

Media coverage and personal experience tend to suggest that rail transport in Britain is unreliable. This negative image is to some extent a consequence of long-term under-investment in the railways, but the situation is exacerbated by rail's inherent operational inflexibility relative to other modes. This characteristic means that rail transport is particularly vulnerable to disruptive incidents, which can cause rapid and widespread disruption and delay to train services. There are two broad, complementary responses to this situation: (i) investment in capacity and reliability, which is long-term and expensive, and (ii) the development and implementation of improved responses to disruptive events. A fundamental measure of rail's performance is the delay incurred by trains, passengers and freight, and the minimisation of delay is one of the major goals of disruption management and of the regulation of trains when disruptions occur.

In order to prevent and measure train delay, accurate journey time information is required. In order to reduce, and, ideally, minimise delay when disruptive events occur, it is useful to be able to simulate a range of possible responses. Such techniques also have wider applicability in railway operations planning. Following a review of the underlying issues, this thesis describes the development of two computer models to address these requirements. These models are also of use to Arup, the Industrial Sponsor of the research activity.

The need for improved methods of train regulation has been acknowledged within the British railway industry. Existing methods have some significant shortcomings, particularly with regard to the current system of train classification and the issue of regulating multiple trains. An alternative, improved classification system is proposed, together with the application of recognised scheduling techniques to multiple train regulation, and their potential benefits are demonstrated.

*The railway is an industry where  
chaos theory applies. If a driver  
is sick in Aberdeen, Birmingham  
New Street falls apart.*

*Modern Railways, January 2002*

## LIST OF CONTENTS

<b>List of Tables</b>	vi
<b>List of Figures</b>	vii
<b>Declaration of Authorship</b>	viii
<b>Acknowledgments</b>	ix
<b>Chapter 1: Introduction</b>	1
Background	1
Objectives	1
Research Approach	2
Thesis Structure	2
<b>Chapter 2: The Disruption of Railway Operations</b>	4
Transport Systems and their Vulnerability to Disruption	4
Railways: History, Background and Characteristics	7
The Particular Vulnerability of Railway Services to Disruption	10
Dealing with Disruptive Incidents: Prevention and/or Cure?	22
<b>Chapter 3: The Development of a Spreadsheet Model for the Calculation of Train Journey Times</b>	30
Origins and Specification of the Model	31
Initial Development of the Excel-Based Model	35
Subsequent Model Development	36
Input to the Model	37
Model Output	39
Segment Boundary Issues	46
Variable Rates of Acceleration and Deceleration	48
Model Validation	54
Future Model Development	56
<b>Chapter 4: The Development of a General Model for the Simulation of Railway Operations</b>	59
Origins and Specification of the Model	60
Track Network Description and Definition	62
Signalling System Description and Definition	64
Timetable Description and Definition	68
Reading and Preparing the Model Data	69
Running the Model	78
Further Work	86

<b>Chapter 5: Disruption Management</b>	88
Capacity, Disruption and Delay	88
Responses to Disruptive Incidents	96
Basic Disruption Management ‘Tools’	98
Disruption Management on Britain’s Mainline Railways	100
Proposed Framework for Disruption Management	111
Further Work	126
<b>Chapter 6: Conclusions</b>	128
<b>References</b>	132
<b>Appendices</b>	
Appendix A: Extracts from Original Arup RUNTIME Documentation	
Appendix B: Results of Train Performance Calculations by W.W. Hay	

## LIST OF TABLES

Table 2.1: Rail Freight's Market Share in Different Locations	9
Table 2.2: Proposed Definitions for Abnormal Working	15
Table 4.1: Data Used to Define Track Network	63
Table 4.2: Data Used to Define Signalling System	66
Table 4.3: Data Used to Specify the Modelled Timetable	68
Table 5.1: Train Classifications in British Railway Operations	102
Table 5.2: Train Data	113
Table 5.3: Regulation Outcomes	114
Table 5.4: Train Travel Times	116
Table 5.5: Delay Resulting From Strict Regulation By Margin Table	118
Table 5.6: Delay Resulting From Re-Sequencing Trains (1)	118
Table 5.7: Delay Resulting From Re-Sequencing Trains (2)	119
Table 5.8: Delay Resulting From Re-Sequencing Trains (3)	119
Table 5.9: Delay Resulting From Re-Sequencing Trains (4)	119
Table 5.10: Job Weights and Processing Times	125

## LIST OF FIGURES

Figure 2.1: Rail's Absolute Share of the UK Passenger Market	8
Figure 2.2: Rail's Percentage Share of the UK Passenger Market	8
Figure 3.1: RUNTIME Model 'Input and Results' Worksheet (Input Data)	37
Figure 3.2(a): RUNTIME Model Summary Tabular Output	40
Figure 3.2(b): RUNTIME Model Summary Tabular Output	40
Figure 3.3: RUNTIME Model Intermediate Journey Times	42
Figure 3.4: 'Maximum and Achieved Line Speeds' Graph	44
Figure 3.5: 'Cumulative Distance vs. Time' Graph	44
Figure 3.6: 'Maximum and Achieved Line Speeds vs. Elapsed Distance' Graph	45
Figure 3.7: 'Maximum and Achieved Line Speeds vs. Elapsed Time' Graph	45
Figure 3.8: Extract from Second-by-Second Output Data	46
Figure 3.9: The Constraint of Acceleration by Preceding Segment Speed Limits	47
Figure 3.10: Tractive Effort and Train Resistance vs. Speed	51
Figure 3.11: Data Input for Variable Acceleration and Deceleration Rates	53
Figure 3.12: Example of Model Output – Speed vs. Distance	54
Figure 3.13: Example of Model Output – Speed vs. Time	55
Figure 3.14: RUNTIME Animation Facility	56
Figure 4.1: Network Data Entry Window	69
Figure 4.2: Signalling Data Entry Window	71
Figure 4.3: Timetable Data Entry Window	73
Figure 4.4: Simulation Preparation Window	76
Figure 4.5: Initial Simulation Window	77
Figure 4.6: Running Simulation Window	78
Figure 4.7: Paused Simulation Window	79
Figure 4.8: Drawing of Trains on Screen	82
Figure 5.1: Train Graphs for Trains of Equal Performance	90
Figure 5.2: Train Graphs for Trains of Different Performance	91
Figure 5.3: Reactionary Delay vs. CUI	93
Figure 5.4: Net Benefits of Additional Train Services	94
Figure 5.5: Train Regulation Example 1	113
Figure 5.6: Train Regulation Example 2	116

**DECLARATION OF AUTHORSHIP**


I, JOHN ARMSTRONG,

declare that the thesis entitled

RAIL JOURNEY RECOVERY FOLLOWING AN  
INCIDENT

and the work presented in it are my own. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:  .....

Date: 11-2-2005 .....



## ACKNOWLEDGEMENTS

At an organisational level, I am grateful to EPSRC for setting up and funding the Engineering Doctorate programme; to the University of Southampton for introducing the ‘Transport Knowledge and Systems Engineering’ theme; and to Arup for supporting the scheme in general and my participation in particular, allowing me to take Leave of Absence, and providing funding for and supervision of my research.

At an individual level, I am grateful to the following people:

In Arup, Tony Kerr and Nick O’Riordan provided early support and encouragement, Ed Humphreys acted as my Industrial Supervisor, and Richard Foster, Stefan Sanders and others provided further advice, guidance and support.

At the University of Southampton, Professor Mike McDonald and Professor Peter Henderson provided me with academic supervision and moral and material support, and Dr Steve Crouch and Dr Bob Walters gave me further assistance with the software development aspects of my work. Anne Donohue was extremely helpful during the period in which my application for the EngD was being processed and subsequently, and Professor Chris Clayton and Dr Neil Ross directed the EngD programme at Southampton during my participation. Melanie Hallford, Davina Channon and Maggie Bond gave me frequent and valuable administrative assistance, and Lance Draper, Ben Waterson and Grant MacKinnon provided me with the necessary computer support.

I am also grateful to Derek Holmes and Richard May at Network Rail, to Alastair Fyfe at South West Trains, to Robert Watson at Robert Watson Associates, to Tom Castor at Laing Rail, and to Chris Bouch at Rail Research UK for organising meetings with various individuals from Network Rail and the Strategic Rail Authority.

## **1.0 INTRODUCTION**

### **Background**

- 1.0 In recent years, media coverage and individual experience have tended to indicate that rail transport is unreliable, and vulnerable to a wide variety of events and occurrences, including infrastructure and train failures, human error, leaf fall and weather conditions, the latter including excessive heat, cold, and, perhaps most notoriously, “the wrong kind of snow”.
- 1.1 The author’s interest in the research topic was stimulated by personal experience of rail travel difficulties, and by the coverage of such problems in the broadcast, print and technical media, both prior to and during his employment as a Transportation Engineer with Arup. The introduction of a transport-oriented Engineering Doctorate (EngD) programme at the University of Southampton provided an opportunity to pursue this interest further, and was consistent with Arup’s growing involvement in rail transport consultancy. The author took leave of absence from Arup for the duration of his EngD, but maintained contact with the company through his Industrial Supervisor and others, agreeing and refining the research objectives, contributing to various related projects, and receiving feedback on the ongoing research effort.

### **Objectives**

- 1.2 The original objective of the research was the investigation and development of techniques for reducing the impact of disruptive incidents on railway operations. These activities require some means of simulating normal and disrupted railway operations, and of measuring the impacts of disruptive events and the effectiveness of possible responses.
- 1.3 In addition to providing the means of addressing the original research objective, the development of simulation tools was of direct interest to Arup, since it had the potential to contribute to the firm’s consultancy activities, in addition to enhancing the knowledge base within the firm. The development of such tools is highly consistent with the ‘product-oriented’ nature of the EngD programme, and became a core objective and ‘deliverable’ of the research activity. The

research goals thus evolved from the original, single objective, to three objectives to be pursued in parallel, as described below.

## **Research Approach**

- 1.4 The objectives evolved into three main strands of research, identified and pursued with input from the Industrial Sponsor. These were:
- (i) the enhancement of an existing Arup in-house spreadsheet model for the calculation of train journey times;
  - (ii) the development of a general model for the simulation of railway operations; and
  - (iii) a review of existing disruption management techniques, and the identification of possible alternatives and/or enhancements.
- 1.5 These three research strands are linked by the common theme of delay. A common measure of the consequences of disruptive events, and therefore of the effectiveness of remedial responses, is the total train and/or passenger/freight delay that occurs as a result. The determination of delay requires that theoretical/scheduled journey times are known or can be calculated: the spreadsheet model provides a means of doing this. In addition to enabling the simulation of railway operations, the general model provides a means of identifying and calculating any delay incurred by simulated trains relative to their journey schedules. Finally, the third strand of the research reviews existing approaches to dealing with disruptive incidents that cause delay, and considers possible enhancements to these, with the aim of reducing delay.

## **Thesis Structure**

- 1.6 Following this Introduction, Chapter 2 considers the sensitivity of all transport systems to disruptive incidents, and compares rail with other modes, identifying the factors that render railway operations particularly vulnerable to disruption. It then summarises the underlying principles of railway operations and the broad categories of response available for dealing with disruptive incidents, and

describes the details and some shortcomings of current methods. Chapter 3 describes the development of an enhanced spreadsheet-based model for the calculation of train journey times, including the origins of the model, the objectives for its improvement, the principles underlying these improvements, and the means by which they were achieved. Chapter 4 then describes the development of a general model for simulating railway operations, including the agreed objectives and aspirations for the model, the means by which these were achieved, and possible avenues for future development. Chapter 5 reviews existing approaches to and methods for the management of disruptive incidents, and proposes an alternative approach. Chapter 6 presents the conclusions drawn from the work described in the preceding chapters, and is followed by a list of References and by the Appendices.

- 1.7 A separate volume, containing two Technical Appendices, was prepared, describing in detail the development and underlying algorithms of the spreadsheet and general models. These documents are confidential, and were made available only for the purpose of assessing the EngD; they are therefore not included with this thesis. The printed copies of the source code for the two models, referred to in the Appendix texts, were omitted from the copies submitted for the EngD assessment, again for reasons of confidentiality.

## **2.0 THE DISRUPTION OF RAILWAY OPERATIONS**

2.1 This Chapter considers the particular vulnerability of railway operations to disruptive incidents. It first considers transport systems in general, our reliance upon them, and variations between modes in terms of (i) vulnerability to disruption, and (ii) user attitudes and reactions to such occurrences. It then summarises the relevant history and characteristics of rail transport, and identifies the mode's comparative vulnerabilities and advantages relative to other modes. The particular vulnerability of railway operations to disruptive incidents is next considered, in terms of user perceptions, system characteristics and operating principles. Finally, two general responses to operational disruptions are identified and discussed, together with their respective advantages and drawbacks.

### **Transport Systems and their Vulnerability to Disruption**

2.2 Modern transport facilities in the developed world enable the relatively cheap, fast, efficient, reliable and safe movement of people and goods. Large volumes of passengers and freight can be carried at high standards of comfort and reliability, over distances and at speeds that would mostly have been impossible until the 20<sup>th</sup> century, and would have been unimaginable prior to the “transport revolution” that has occurred since the late 18<sup>th</sup> century (Bagwell, 1974, p11). These feats are achieved by means of complex systems of infrastructure, vehicles and operating practices, based on decades (and, sometimes, centuries) of investment, development and experience; as Dalla Chiara and Gačanin (2004, p6) observe, “rail can boast a tradition dating to the first half of the 19<sup>th</sup> century.”

2.3 The scope of these systems, the reliance upon them (Bagwell and Lyth, 2002, p. xiii) and the extent to which they are taken for granted (ibid, p. xi), and the sheer scale of the tasks they perform are perhaps most graphically illustrated by their occasional failures, and the consequences for their users in terms of congestion, delay and the possible loss of or injury/damage to passengers and freight. Users of road, rail and air transport networks are often all too familiar with a certain level of ‘background’ congestion and delay which results from

transport networks operating at or close to their capacities. The water-borne modes are perhaps less vulnerable to these problems, being generally less widely and frequently used by passengers than other modes, and having a certain ‘infrastructural redundancy’ on large bodies of water. Nonetheless, services may be disrupted, for example when heavily-trafficked rivers like the Rhine, Danube and Mississippi are affected by flood or drought (National Safety Council, 2000; BBC News, 2003), and when industrial action in US Pacific ports in 2002 effectively closed them to container traffic (BBC News, 2002).

- 2.4 For most transport users, however, travel disruption is more likely to be experienced on the roads, in the air (or, more likely, at airports) and on the railways. The everyday, ‘background’ levels of congestion and delay noted above can be greatly exacerbated by disruptive incidents such as road accidents, air traffic control problems and rolling stock, track or signalling failures. Transport facilities which normally operate at a high capacity, such as multi-lane highways, major airports and important rail routes, are particularly vulnerable to the effects of such incidents, handling as they do large volumes of traffic, whereby large numbers of travellers and volumes of freight may quickly be affected, with widespread knock-on effects on upstream transport links and downstream connections.
- 2.5 Everyday congestion and delay is a common feature of road transport, despite (and, arguably, because of) the development and provision of extensive, high-capacity highway networks. Peak-period commuters to and from major centres, and, increasingly, off-peak, inter-city and other road users are routinely subject to these problems. However, the users of private road vehicles are travelling by a mode and route, and at a time, largely of their own choosing, and they contribute to any delays experienced by themselves and other road users. Even when delayed, they can still ‘enjoy’ the comfort and comparative privacy of their own vehicles. These ‘routine’ delays can be greatly extended by an accident on a motorway, for example, but the same mitigating factors still largely apply. Under normal, and even extreme conditions of delay, road users typically enjoy at least some limited flexibility in terms of choice of alternative

routes, and existing and emerging in-vehicle technology enables road users to avoid and/or more effectively respond to congested road conditions.

- 2.6 Users of road-based public transport (i.e. taxis and buses) may have less say in their choice of mode, and bus users certainly have fewer choices when it comes to choosing their routes and times of travel. Although they are still vulnerable to the general disruption and delay experienced and caused by road traffic (notwithstanding such measures as the introduction of dedicated bus and taxi lanes, etc.), these problems ‘come with the territory’ and there is relatively little the service users and providers can do directly to influence the situation and reduce such problems as arise.
- 2.7 Despite the growth of air travel and the rise of the budget airlines, it remains a comparatively ‘young’ mode, and less of an everyday experience for most people than road or (in much of Europe, anyway) rail travel. Modern aircraft are, of necessity, highly reliable machines, and the majority of delays experienced by users are probably due to air traffic control problems, unscheduled aircraft maintenance, or weather conditions, although the last factor is less of an issue with improved navigational facilities and aircraft technology. These factors, together with the safety-critical nature and residual exoticism of air travel, may make users more tolerant of slight delays and accepting of the knock-on effects of disruption. When disruptive incidents do occur, the mode also enjoys a degree of ‘infrastructural redundancy’, in that incoming aircraft may be ‘stacked’ or diverted to alternative landing sites. A major and obvious drawback of aircraft is that, while airborne, they cannot be ‘parked up’ and are usually totally dependent on limited supplies of onboard fuel. However, this adds an urgency to dealing with disruptive incidents, and requires that a certain adequate level of back-up options and system redundancy be available. The ability of the mode to deal with disruptive incidents was illustrated by the rapid clearing of US airspace by civilian aircraft on September 11, 2001, in conditions of extreme confusion and urgency.

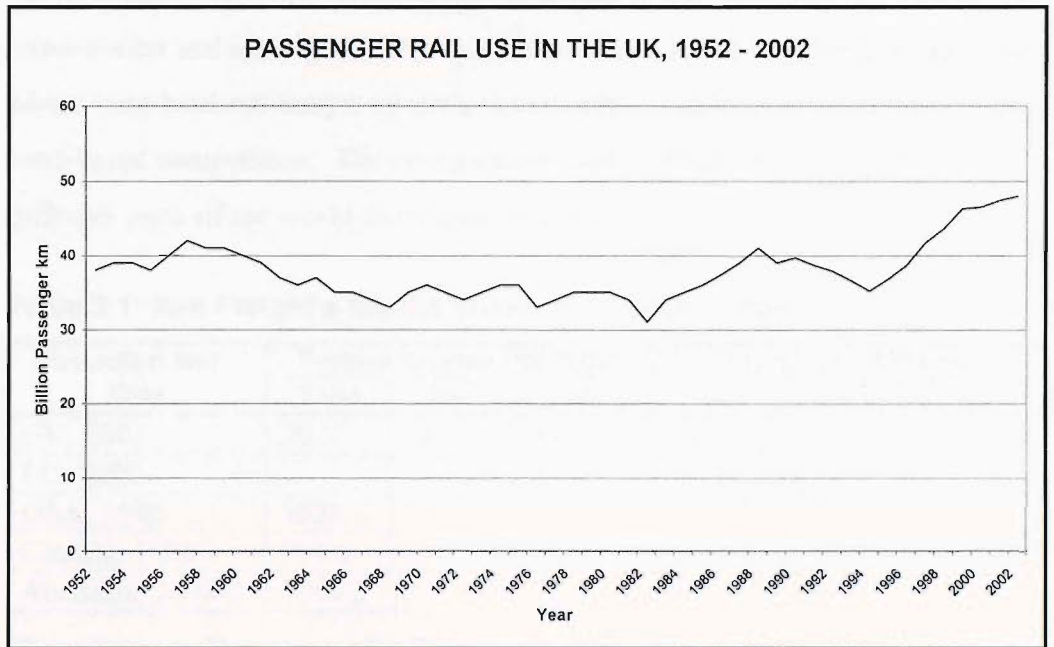
## **Railways: History, Background and Characteristics**

- 2.8 The railway is one of the longest-established of the modern, ‘industrialised’ transport modes. They date in their current form from 1825 (slightly later than the advent of the steamship, although they predate the latter’s widespread use), considerably earlier than the internal combustion engine and the aircraft. Canal-borne and animal-powered turnpike traffic were almost completely superseded by the speed, capacity and rapid spread of the railways. According to Faith (1993, pp7-8) “the railway was the first, the most universal and the most dramatic mechanical intrusion into the lives of people and nations, the first of the technical revolutions which created the world as we know it today.”
- 2.9 With influences extending to the standardisation of time zones within their areas of operation (Faith, 2000, pp3, 16; Goddard, 1994, p14; Schivelbusch, 1986, p43), the railways were among the largest industrial organisations of the 19<sup>th</sup> and early 20<sup>th</sup> centuries, employing enormous resources of staff, infrastructure and rolling stock, widely dispersed over large areas. The efficient and harmonious use of these widely distributed resources was a major challenge, particularly in the absence of today’s communication and computational technologies. Faith (1993, pp62, 67) observes that “the great railway systems, especially in Britain and the United States, were the first modern businesses, relying on increasingly elaborate and pioneering management techniques” and their leaders were “the pioneers of modern industrial management.” The railways pioneered many techniques for the management of resources, making particular use of graphical methods for timetabling and resource allocation.
- 2.10 The invention and development of the internal combustion engine and the self-propelled road vehicle in the late 19<sup>th</sup> and early 20<sup>th</sup> centuries saw a resurgence in road travel, and the development of comprehensive highway networks, including systems of limited-access, high-speed motorway-type routes. These developments, together with the rise of air travel, took large proportions of growing local and long-distance passenger and freight traffic away from the railways, particularly in the second half of the 20<sup>th</sup> century (Wolmar, 2001, pp27-28, 33; de Fontgalland, 1984, p3). Although rail’s absolute share of the

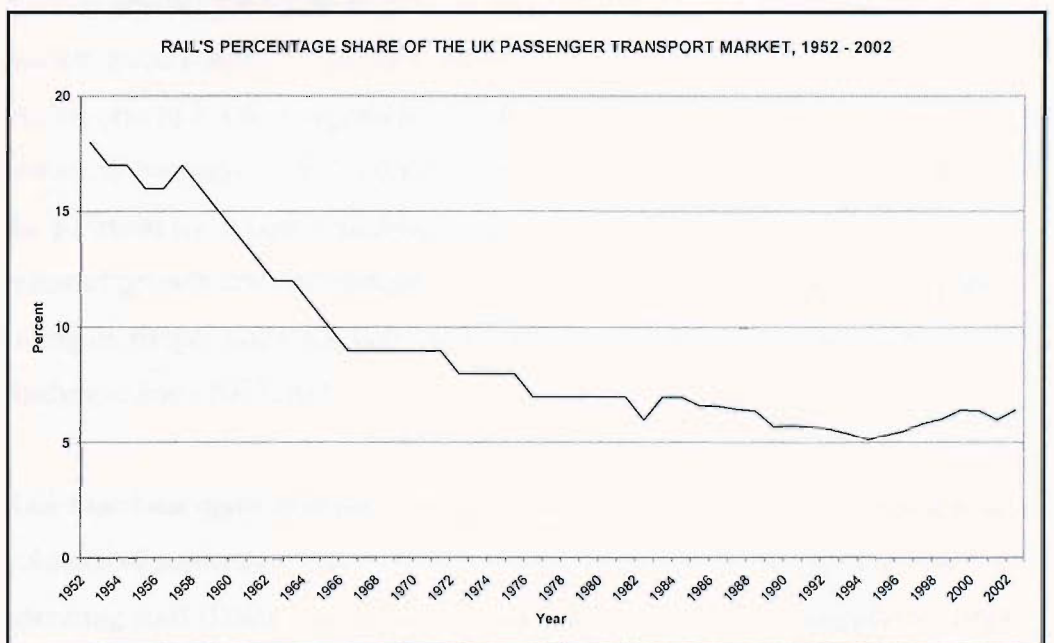


UK passenger transport market remained quite steady during the latter period, at approximately 40 billion passenger km per annum, its percentage share of the market declined from 18% to 6%, (Department for Transport, 2003a). Wolmar (2001, p45) also observes that “the number of train journeys has remained remarkably steady since World War II while car journeys have increased massively.” The DfT data are displayed in Figures 2.1 and 2.2, below.

**Figure 2.1: Rail’s Absolute Share of the UK Passenger Market**



**Figure 2.2: Rail’s Percentage Share of the UK Passenger Market**



2.11 More recently, railways have seen something of a resurgence, as can be seen from Figure 2.1. The introduction of high-speed passenger services on purpose-built, dedicated routes in Japan, Continental Europe (most notably in France) and elsewhere has given rail a considerable journey time advantage over road transport, and, increasingly, for short- and medium-length journeys, over air travel (Goddard, 1994, pp264-265; Vranich, 1991, pp36, 42, 49-51). Rail freight remains highly viable in some sectors, particularly in large geographical entities such as the USA, Canada and Australia, which are relatively free of the cross-border and interoperability constraints found in Europe, for example, and where long-haul rail freight operations can offer considerable advantages over road-based competition. The comparative market shares of rail freight in different parts of the world are shown in Table 2.1.

**Table 2.1: Rail Freight's Market Share in Different Locations**

Location and Year	Tonnes Carried (Millions)			Tonne km (Billions)		
	Total	Rail	% Share	Total	Rail	% Share
UK, 2001	2037	94	4.6	247	19	7.7
EU, 2000	-	-	-	3078	249	8.1
USA, 1996	7320.7	1461.4	20.0	5916.2	1979.7	33.5
Canada, 1996	734.6	200	27.2	614.3	221.4	36.0
Australia, 2001/2	2069.2	535.1	25.9	374.9	137.7	36.7

(Data Sources: Department for Transport, 2003b; Eurostat, 2003; U.S. Census Bureau, 2000; Bureau of Transport and Regional Economics, 2003)

2.12 Even in smaller countries such as Britain, rail freight remains competitive in market sectors such as bulk haul and intermodal transport, as demonstrated by the success of EWS, Freightliner, GB Railfreight and Direct Rail Services in attracting business. EWS' operations through the Channel Tunnel demonstrate the potential for longer-haul freight operations in Europe, too, and have now resumed growth after the disruption and hiatus caused by illegal 'passengers' riding on freight trains through the Tunnel from Fréthun yard (*Modern Railways*, June 2003, p11).

2.13 Rail-based transport systems provide an energy-efficient means of moving large volumes of passengers and freight over long distances, using small numbers of operating staff (Dalla Chiara and Gačanin, 2004, pp6-9; De Fontgalland, 1984,

p9). As noted earlier, the mode has characteristics (speed, city centre to city centre service) which give it a considerable inherent advantage in some market sectors. Increasing congestion and journey time uncertainty on the roads and in the air should also serve to increase the perceived attractiveness (or, at least, decrease the perceived unattractiveness) of the rail mode. For this advantage to be exploited fully, however, it is important that the rail mode's own vulnerability to disruption and unreliability be addressed, and reduced as much as is reasonably possible.

## **The Particular Vulnerability of Railway Services to Disruption**

### User Perceptions

- 2.14 It is argued above that users of the road and air modes of transport may have a certain tolerance of delays and disruption to services. This is on the basis that road-users are often 'part of the problem', while air travel is a comparative novelty for many users, and safety is of such paramount importance in the airline industry that delays and disruption may be considered to be less unacceptable than for other modes. Indeed, Faith (2000, p4) quotes a saying in the airline industry that it is "better to be late in this world than early in the next." The railway industry does not benefit from these mitigating factors, however: Faith (op cit, pp3-4) quotes the railway historian Norman Pattenden as saying that "the expectation of timekeeping on railways is very different with the customer from perhaps other forms of transport" and that "there appears to be a different ethos on timekeeping within the railway industry [in which] respect people may be hypercritical."
- 2.15 While the technologies of signalling and rolling stock have obviously made enormous advances since the advent of the railways, the basic technology of 'steel wheel on steel rail' is largely unchanged from the early 19<sup>th</sup> century (apart, of course, from the replacement of iron with steel). It is understandable, therefore, that the travelling public and freight shippers should feel that the knowledge, techniques and skills necessary for successful railway operations should already be in place. For these reasons, in the words of an editorial in *The*

*Independent* (19 April, 2003), “travelling by rail should not be a journey into the unknown.”

- 2.16 Furthermore, it should clearly be possible for railways to operate in a well-planned, efficient manner, since the industry has complete control over access to its infrastructure, unlike the operators and users of highway networks. This is most obviously the case where the railway infrastructure and rolling stock is owned and operated by a single, typically state-owned, entity, but it also holds in the fragmented UK railway industry, whereby train paths are allocated to Train and Freight Operating Companies in such a way as ideally to ensure that the resulting timetables work efficiently and, under normal circumstances, without conflicting train movements which may cause delay. Dalla Chiara and Gaččanin (2004, p6) include “guided, programmable operation” and “manageable quality of service (speed, frequency, price)” among the major advantages of rail transport, while De Fontgalland (1984, pp8-9) asserts that rail’s ‘guided nature’ “constitutes a considerable [operational] advantage, as compared with other forms of transport, ... mak[ing] it possible to schedule traffic in minute detail.”
- 2.17 There are some circumstances and factors which are beyond the complete control of the relatively ‘self-contained’ railway industry, however. One obvious candidate is the weather, particularly strong winds, heavy rain and – notoriously - certain types of snow. While it is difficult in a temperate country like the UK to take all possible extremes of climate into consideration, other countries seem to be able to maintain railway operations under far more adverse weather conditions, and it would seem to be sensible to design and maintain infrastructure and vehicles to withstand most of the conditions that may be anticipated, and to have contingency plans for the rapid clearing of trees felled by storms, etc.
- 2.18 Another factor which can affect railway operations but is largely beyond the railway industry’s control is the obstruction of railway lines by road vehicles, as happened at Great Heck on the East Coast Main Line on 28 February 2001 and has happened several times since, although thankfully with less serious consequences [note: this thesis was completed and submitted shortly before the

crash at Ufton Nervet in Berkshire on 6 November 2004]. Similar problems can be caused by motorists entering level crossings when the gates have started to close, by ‘bridge strikes’ by road vehicles (BBC News, 2001), or by the malicious placing of debris on railway lines.

- 2.19 There is clearly a limit to what the railway industry can do to eliminate or even reduce the likelihood of such externally-caused incidents, particularly where ‘Acts of God’ are concerned, but the general state and reliability of the railway infrastructure, rolling stock and operating procedures are obviously in its control, within the limits of available funding.
- 2.20 For the various reasons cited above, users of the railways are perhaps less likely than users of other modes to be tolerant of disruptive incidents and associated delays and enforced alterations of travel arrangements. Furthermore, whatever the causes of such incidents, railways have, as noted above and elaborated below, characteristics which increase their susceptibility to disruption.

#### System Characteristics and Operating Principles

- 2.21 A key feature of railways is their inherent inflexibility when compared with other modes. As Armstrong (1998, p125) puts it, “the railroad is classified as a “single degree of freedom” mode of transport”, while De Fontgalland (1984, p7) describes the railway as “function[ing] in one-dimensional space.” White (2003a, p16) summarises the situation as follows:

*Among transportation modes, the railroad is made unique by the tracks. Unlike the operator of any other general-use transportation vehicle, the operator of a train does not steer. Trains can only go exactly where the tracks go. They cannot swerve to avoid a collision, they cannot change routes or even lanes of the same route except where a track arrangement has been constructed for the purpose.*

Compared with other modes, railways thus have an inherent lack of a straightforward ‘overtaking’ facility; i.e. when a track or section of line is

obstructed, the flexibility to operate around the obstruction is dependent upon the number of parallel tracks, the proximity of crossovers, the nature of the signalling system (uni-directional or bi-directional), and the density of opposing and/or parallel traffic. On a wider scale, further constraints include driver knowledge of and acceptance of train types on alternative, 'parallel' routes.

2.22 Another limiting factor is the typical requirement for trains, particularly passenger services, to make intermediate stops at scheduled times between their termini. While these constraints obviously also apply to other public transport services, the relatively low density of rail networks compared with highway systems means that there are typically far fewer routeing alternatives available, and the more closely spaced are the scheduled stopping points, the greater the inherent constraint.

2.23 A recent report by Lloyd's Rail Register (2004, p1) for the Rail Safety and Standards Board (RSSB) lists some of the possible causes of disruption to normal operating conditions on the railway:

- *The introduction of a new timetable;*
- *The breakdown of normal equipment, such as:*
  - *signalling and other control systems such as Train Protection and Warning System (TPWS) and Automatic Train Protection (ATP);*
  - *power supply systems;*
  - *communications systems;*
  - *onboard train systems;*
  - *station control arrangements;*
- *Planned or emergency engineering work;*
- *Temporary or longer-term loss of routes and consequent diversions, such as the blockades on the West Coast Main Line;*
- *Introduction of temporary or emergency speed restrictions;*
- *Bad weather or natural disasters;*

- *Major industrial disputes;*
- *Extreme overcrowding and abnormally large movements of people, for example major holidays and sporting events.*

On the same page, the report makes the following distinctions: “‘normal’, ‘abnormal’ and ‘emergency’ working, and ‘degraded’ modes of operation” and notes that previous RSSB research “found that there is considerable variation in the [terms’] understanding and interpretation ... across the rail industry.” It therefore sought to determine

*consistent definitions ... as the first step in developing a ‘best practice’ policy to provide a defensible system that maximises the opportunities for controlling and monitoring abnormal and degraded working.*

2.24 In the course of the Lloyd’s Rail Register research, workshops were held with representatives of the rail industry, in the course of which the following definitions were proposed (Lloyd’s Rail Register, 2004, p17):

- |                                 |   |
|---------------------------------|---|
| <b><i>Normal working:</i></b>   | <i>This describes trains running under normal signals up to permissible speeds.</i>   |
| <b><i>Amended working:</i></b>  | <i>This describes trains running under normal signals but not at permissible speeds or not on the normal route.</i>                                     |
| <b><i>Degraded working:</i></b> | <i>This describes trains running when the normal signalling arrangements are not in place for whatever reason.</i>                                      |
| <b><i>Abnormal working:</i></b> | <i>This describes the situation where trains are running normally, but there is a hidden defect that threatens the safety of the railway operation.</i> |

**Emergency working:** *A[n] unforeseen or unplanned event which has life threatening or extreme loss implications and requires immediate attention. This is unchanged from the RSPG [Railway Safety Principles and Guidance] definition.*

These definitions are elaborated upon in Table 5 of the report (Lloyd’s Rail Register, 2004, p18), reproduced as Table 2.2, below.

**Table 2.2: Proposed Definitions for Abnormal Working**

Type of Working	Category	Impact on Train Operations	Cause/Condition (examples)
Normal	1	Trains running under normal signals at permissible speed.	Train and infrastructure within maintenance tolerances.
	2	Train running under normal signals at permissible speed but with possible reduction in passenger comfort or controlled increase in risk.	Part of train or infrastructure below maintenance tolerances but above safety limits and within repair timescales.
Amended	3	Train running under normal signals but with a published reduction in speed or deviation from normal route.	Application of a Temporary Speed Restriction. Need to apply a change of platform. Introduction of new timetable.
	4	Train running under normal signals but with an unpublished reduction in speed or deviation from normal route.	Part of train or infrastructure below maintenance tolerances and outside repair timescales but kept within safety limits by mitigation. On-train defect requiring speed reduction. Application of an Emergency Speed Restriction. Examination of the line (under clear signals). Passing train over broken rail. Train diverted from Fast line to Slow line or vice versa.



Type of Working	Category	Impact on Train Operations	Cause/Condition (examples)
Degraded	5	Train not running under normal signals or signalling arrangements.	Signal equipment failure/disconnection. Single Line Working. Examination of the line (passing a signal at danger). Movements to, from and within possessions. Track circuit failure. Level crossing failure. Failure of block signalling equipment.
Abnormal	6	Train running under normal signals at permissible speed but at risk of serious incident.	Part of train or infrastructure outside maintenance tolerances/safety limits and with no mitigation in place. Serious technical defect undiscovered.
Emergency	7	Immediate emergency action required to prevent a catastrophic accident	Derailed train obstructing adjacent line. Road vehicle coming to rest on line. Train running away (SPAD).

As would be expected from a report for the RSSB, the emphasis is on the safety implications of the various types of working. However, the classification is also useful in the current context, for clarifying and categorising the broad types of disruptive incident that may occur; the amended, degraded and emergency categories are of particular interest with respect to operational disruptions.

2.25 When disruptive incidents do occur, the comparatively restricted range of possible responses, described above, is compounded by the rate at which the knock-on effects can spread through affected parts of the network. According to Network Rail (2003, p5),

*with many parts of the network operating ever closer to full capacity, the railway is highly sensitive to even minor sources of disruption. Whether the initial cause of disruption is an individual delay to a single train or an infrastructure fault, the knock-on consequences rapidly multiply and can result in*

*significant disruption to train services in the area for some hours afterwards. This results in an increase in the overall delay per incident, which lies behind a substantial part of the deterioration in performance since 1999/2000. While the numbers of failures of signalling and other non-track assets have remained broadly unchanged, the total delay caused has increased substantially.*

Such effects can also spread widely: as Jack (2001, p65) observes, “one delayed train can cause other delayed trains for hundreds of miles down the track.” This characteristic was illustrated by the derailment of a freight train at Quintinshill on the West Coast Main Line in Scotland on June 17 2002, when the following day’s edition of *The Independent* reported that “the knock-on effects for the network were so severe that disruption extended as far as the West Country [of England].”

- 2.26 In order to allow the safe operation of any vehicular transport system, it is obviously essential to maintain sufficient stopping distances between successive moving vehicles. Because of trains’ typically high weights and speeds, and the relatively low coefficient of friction between steel wheel and steel rail, their stopping distances are much greater than for road vehicles (Dalla Chiara and Gačanin, 2004, p16), and they thus need to be kept farther apart. This necessary separation is maintained by means of signalling systems. In the UK, these typically take the form of ‘fixed block’ systems, whereby the track is divided into successive linear sections called blocks, which cannot usually be occupied by more than one train at a time. This means that when a series of trains have to stop behind each other, they are quite widely separated, and any blockage can extend ‘upstream’ quickly and over a long distance. The potential consequence of this is that, as congestion spreads upstream of a site of disruption, possibly extending across junctions and through stations, even trains which are not using the directly affected section of the network may be subjected to delay. This characteristic can be mitigated by means of ‘moving block’ signalling systems, as used on the Docklands Light Railway and elsewhere, whereby the minimum spacing between trains is reduced to the required braking distance(s) of following train(s), with suitable additional distance allowance(s) added for

safety purposes. Trains can typically thus stop much closer together than is the case with fixed block systems, reducing the potential lengths of queues.

2.27 Given these inherent characteristics of the railways, and the relative ‘handicaps’ they thus face, it is all the more important that they be operated as efficiently and reliably as possible. Glover (1999, pp11-18) provides the following list of 12 principles of operation:

1. *The service which railway companies provide and for which they are paid is movement;*
2. *Faster journey times allow the same rolling stock and staff to provide a more frequent service;*
3. *Load factors are all-important;*
4. *Peak demand is difficult to cater for economically;*
5. *Rail traffics are interdependent;*
6. *Line capacity is a scarce resource;*
7. *Short turnarounds are a key to utilisation;*
8. *Good performance is vital;*
9. *Surplus facilities may be an embarrassment;*
10. *The customer is the best judge of what he wants;*
11. *The railway does not exist in a vacuum;*
12. *Change takes time.*

2.28 Reflecting the earlier observations about rail’s tradition dating back to the 19<sup>th</sup> century, and the long-standing nature of the industry’s basic technology, these principles are neither radical nor new. Samuel (1961, pp9-21) lists ten “simple common-sense” principles of efficient railway operation. These are:

1. *Keep traffic and mobile equipment moving;*
2. *Speed is economical to the operator;*

3. *Operating units are interdependent and their reaction on each other must always be considered;*
4. *The advantage of rail transport over road transport can best be exploited over long distances with capacity loads;*
5. *The potential capacity of operating units is not fixed, but is dependent on the method of using them;*
6. *The retention of operating units in excess of needs involves unnecessary costs and leads to inefficient working;*
7. *Peak demands militate against the full use of equipment or of staff and may lead to uneconomic results;*
8. *Realistic planning ahead gives the best results, but current modification is often necessary to meet the situation;*
9. *Reliability is vital;*
10. *Improved operating methods may result in operating units becoming redundant.*

These correspond quite closely to Glover’s list, confirming that the basic principles involved have not changed much in the almost 40 years separating the publication of the two lists, and again highlighting how relatively well-established and long-standing these principles are – this is borne out further by reference to still earlier works, some of which are quoted below.

- 2.29 Of Glover’s list, principles 1, 5, 6, 8 and 9 are probably of most relevance here. Under principle 5, he (p13) observes that “if everything proceeds according to plan and to timetable, all will work as intended [but] the system must be robust enough to cope with change [since] there will always be some instances of equipment failure, staff not turning up on time, holding trains for late-running connections, and so on.” He therefore concludes that “the system of operation must be resilient”, but under principle 9, (pp16, 17), while acknowledging that “some spare capacity allows flexibility”, concludes that “truly spare capacity, be it in infrastructure, rolling stock or facilities, represents a cost to be avoided wherever possible.”

2.30 In Samuel's list, principles 1, 3, 8 and 9 are particularly relevant in the current context. Under the heading of the first principle, Samuel observes (p11) that "congestion must be avoided ... in order to maintain movement", and describes it as "one of the deadliest enemies of efficient operation, because its effect tends to spread outwards from the trouble spot", resulting in ever-widening paralysis. He describes the avoidance of congestion as "one of the main functions of the Control organisation". Under the third principle he acknowledges the "railway's inherent disadvantage of being track-bound", and how this inflexibility means a late-running train can affect following, preceding, connecting and conflicting trains, so that the consequences "may be felt hundreds of miles away", echoing the example quoted and the observations made above. Obviously, the more intensively a railway network is used, the more rapidly and widely these effects will spread. Under principle no. 8, he acknowledges the importance of planning, but also the need to modify plans in response to unforeseen circumstances, and describes such modifications as the "purpose of the Control organisation". Finally, under the ninth principle, he observes that "if reliability is vital to the operator it is imperative to the customer" and identifies unreliability as "one of the main factors which has caused the transfer of traffic from rail to road".

2.31 Expanding on the inherent inflexibility of railways, Samuel (pp24, 25) identifies two particular handicaps: the effect of obstructions and the 'overtaking problem'. He lists the three possible responses to an obstruction as being to:

1. *divert trains by alternative routes;*
2. *on double or multiple track, run trains in both directions on the unobstructed line(s), although this is slow and tedious; or*
3. *cancel or postpone movements on the affected line until the obstruction is cleared.*

He acknowledges that all the above responses will inevitably result in delay, and contrasts the effects with those resulting from a (minor) obstruction on a highway, which can be passed with comparative ease. On the railway, the

overtaking of slow or stationary trains (or other track obstructions) is only possible where two or more tracks are provided in one direction, passing loops or sidings are provided, or two-way running is possible on an adjacent track. However, all of these options are “expensive in provision and maintenance of track and/or signalling.” On page 41, he summarises these issues by saying that, where possible, obstructions should be prevented, and other requirements for overtaking should be avoided, but that “their effects can be minimised by the maximum flexibility of alternative routes and tracks.” However, he notes (p164) that “train crews can only work unassisted over those routes where they are thoroughly acquainted with the signalling lay-out. This is referred to as route or road knowledge.”

- 2.32 Since Samuel cautions elsewhere (p22) that “the object is to run the required train service with the minimum land, the fewest tracks and at the lowest total cost for track and signalling facilities”, there are obviously conflicts between providing maximum flexibility and minimising infrastructure costs. This echoes Glover’s comments quoted in para. 2.29, and highlights the importance of making the best and ‘cleverest’ possible use of the resources available. In specific reference to station operations, Hare (1927, p17) also notes the “desirability of increasing ... capacity ... by improved methods of working, rather than by the provision of extra accommodation.” However, the consequences of eliminating spare capacity are summarised by Schmid (2003, p11-12):

*Although originally very large, by the 1980s Britain’s railway infrastructure had been pared back to the essential, both in terms of route length and the provision of diversionary facilities, sidings and terminal connections. Whilst this was attractive from a management as well as a maintenance volume point of view, it created risks for service stability. Very limited numbers of passing loops and long single track sections (e.g. the branch line from Burnley to Colne) no longer allow the recovery from disruptions in train movements and make the re-introduction of freight services almost impossible on some routes. Access for*

*maintenance is a significant problem because of the lack of diversionary routes.*

The role of junctions, crossovers and bi-directional signalling in facilitating operational flexibility is also noted by the Institution of Railway Operators (2004, pp30, 31, 34).

### **Dealing with Disruptive Incidents: Prevention and/or Cure?**

#### Prevention: Improved Reliability and Provision of Redundancy - Investment

- 2.33 One approach to dealing with disruptive incidents is to seek to prevent them happening in the first place, by investing in network capacity and in vehicle and infrastructure reliability. Investment in reliability reduces the frequency of disruptive incidents, while investment in capacity provides some ‘slack’ in the system, reducing the rate at which the knock-on effects of disruptive incidents can spread through a network, and also providing some redundancy and alternative routeing options when responding to such incidents.
- 2.34 Such investment is certainly needed in the UK, both to address a long-standing maintenance and investment backlog, and to provide the necessary capacity and reliability to accommodate increasing volumes of rail traffic. However, these are long-term goals, and require enormous investment. Also, as noted above, the provision of excessive capacity may be wasteful, particularly when other projects and other sectors of the national infrastructure and wider economy are competing for investment. Ironically, the introduction of system upgrades can itself be a source of disruption, as illustrated by the recent West Coast Main Line blockades, and by the problems that are sometimes caused by over-running track ‘possessions’.
- 2.35 While the necessary system improvements are being introduced, and generally, an advantageous and complementary approach is to respond to disruptive incidents in such a way as to minimise the resulting problems. This is the topic of the following section.

Cure: 'Making the Best of a Bad Lot' – Responding Cleverly

2.36 Signalling and Control are of vital importance to the efficient operation of railways, both in the course of normal operations and when disruption occurs. As Hall (2001, p7) observes, “if trains always ran on time and never broke down there would be little need for a signalling system [and] the timetable ... could be devised [so] trains would always be a safe distance apart.” While this approach worked in the early days of railways (and, with modifications, is still in use on some lightly-used freight lines in the US, for example), increased speeds and frequencies of trains required the introduction of signals. According to Lamb (1941, p168), “whilst signalling was originally introduced with a view to safety, it now has the added purpose of keeping trains moving [and] modern signalling implies the control of the trains in such a way that the utmost possible use is obtained from the track.” In a similar vein, Hare (1927, p8) observes that “signals, though still used to stop trains where this is absolutely necessary, are, or should be, arranged so as to enable the maximum necessary number of trains to be kept moving.” This is in keeping with his statement elsewhere (c1931, p3) that “the aim of railway operation may be said to be to find means of keeping traffic moving between its points of origin and destination”, and with both Glover’s and Samuel’s first principles.

2.37 While signallers are primarily concerned with the running of individual trains, the role of the Control organisation relates more to the ‘big picture’. Samuel (1961, p169) described a Control office as “a focal point from which operations are directed.” According to Lamb, writing in 1941 (p207),

*[Traffic Control] results in the more punctual running and speedier working of trains, the better loading of trains, an increase in the train miles per engine hour, and a reduction in light running and empty haulage. Moreover, line capacity has thereby been increased and the duties of train men arranged to better advantage.*

He goes on to say (p209) that



*The principal advantage of train control lies in the concentration of supervision throughout a given area, and the consequent ability to arrange the working in that area, hour by hour, according to the prevailing circumstances. Thus, instructions can be given promptly for the cancelling of trains in case of insufficient traffic, or the running of specials to meet an unforeseen surplus. In the event of congestion, arrangements can at once be made to divert freight traffic by alternative routes to the less affected yards and sidings.*

The importance of control in relation to the management of disruption can thus be seen.

2.38 Nock (1966) reinforces this by observing (pp193, 194) that

*the detailed control of all train movements from a central point ... becomes more and more important as the volume of traffic approaches the capacity of the system to carry it [with the benefit that] the Controller is able to give instructions ... in order to minimize, as far as possible, the effects of delays due to bad running, or to congestion due to variations in the volume of traffic.*

2.39 There is a close relationship between the two functions of signalling and Control, and, with the advent of modern signalling technology, considerable overlap between them. At the time Samuel was writing (1961), he identified (p178) “train regulation [as being] primarily the responsibility of signalmen” and observed that “there are times when the signalmen may ask for assistance from the Control [, which] is aware of the situation over a much wider field.” He also noted that at times “it is necessary for Control to give definite instructions to signalmen” in regard to the prioritisation of trains.

2.40 More recently, however, Hall (2001, p28) describes the signaller's role as being

*to set the routes for trains, and clear the signals, in accordance with the timetable, and when trains run late it is his job to minimise the effect of such late running. He therefore needs to know about trains approaching his area of control so that he can make the most appropriate regulating decisions, and he can see where all trains are by looking at the control panel [of a modern power signalbox].*

This suggests that at least some of the role of Control has been devolved to the individual signallers (who individually cover much larger areas of the railway network than before), although Glover (1999, p36) notes that the widespread introduction of power signalboxes after 1955, with their displays covering large sections of the railway network, meant that “the progress of trains can be scrutinised from a position behind the signalmen, which enables the scene as a whole to be monitored and alternative courses of action to be determined where difficulties arise.”

2.41 Whatever the exact demarcation of responsibilities, Glover (1999, p96) notes that

*the key objective [of the Control organisation still] is to ensure that the planned timetable service is operated punctually and efficiently. In the event of an unplanned incident, [it] is responsible for restoring the service in the most expeditious manner, liaising with the operating companies and any outside organisation which might be involved.*

2.42 Modern, centralised signalling and Control facilities offer great advantages in the event of disruption compared with their predecessors, when the large number of signalboxes, difficulties of communication and resulting incomplete information made it difficult to “make the best decisions regarding the priority to be afforded to any particular [train] movement [and led to] unnecessary delay

to trains on the one hand and wasted line or platform capacity on the other” (Hall, 2001, p9).

- 2.43 Irrespective of the technology available and the working methods used, it remains the case that “when there is divergence from the timings laid down, it is necessary to minimize the reactions and to see that any delay which is unavoidable falls on those trains which can bear it best” (Samuel, 1961, p175). In other words, delay and disruption should be minimised, and every effort should be made to allocate delay in such a way that the most important trains are delayed least. He notes elsewhere (p228) that “careful attention [should be given to] the convenience of passengers. It may be necessary to give more weight to majorities than to minorities, or to long-distance passengers over casual short-distance passengers, etc.” Hare (1927, p71) supports this view, saying that “the general aim must be to provide the maximum of convenience to the majority of passengers, giving rather special consideration to those who are making the longer journeys.” By ‘weighting’ trains according to their importance, and multiplying the delay incurred by each train by its weight, the objective should be to minimise the total weighted delay. However, neither the weighting nor the minimisation is necessarily easily achieved, particularly when many services of different types are affected by a disruptive incident. As Samuel says (p179), normally, “signalmen must regulate according to the priority classification (headcodes) of the trains concerned. But there are occasions when a train of low classification needs priority over one of high classification.”
- 2.44 As an illustration of the potential complexity of the decisions involved, Samuel (p195) stresses the importance of “ensuring that the lower classified trains are not held unnecessarily” and cautions that “undue emphasis on the punctuality of the more important trains may result in less important ones being detained when in fact there was an opportunity for them to run without causing delay to other services.” He is referring specifically to freight trains in this regard, but similar considerations apply to passenger services. Indeed, elsewhere (p206) he observes that when disruption occurs, “the priority between passenger and freight services has to be balanced.”

2.45 The power signalboxes referred to above are in turn being superseded technically by Integrated Electronic Control Centres (IECCs). According to Hall (2001, p11) these are enabled

*by changes in traffic patterns and technology. The pattern of train services today is much more stable and repetitive than previously and certainly more predictable. Passenger train timetables are now generally based on an even interval, the pattern being repeated each hour. Cancellations, special trains and other deviations from the plan are relatively infrequent. Freight trains are not frequently seen on many routes.*

The extent of automation employed means that “an IECC signaller ... can concentrate on his train regulating decisions without the heavy and physically tiring work of the signaller in a manual signalbox” (Hall, 2001, p12).

2.46 Hall goes on to say (p31) that

*a predictable and repetitive train service lends itself readily to the use of computer-controlled automatic route-setting [ARS] equipment. [In IECCs] routes are set by the computer as programmed, including decisions on priorities at junctions, and alterations in the case of late running.*

2.47 However, this raises a question as to what happens in the event of severe disruptions, particularly if the automation is based on the assumptions of stable, repetitive and predictable services, and “infrequent ... deviations from the plan.” The assumed infrequency of freight services also seems to be contrary to government transport policy, as expressed in the 1998 Transport White Paper (Department of the Environment, Transport and the Regions, 1998) and elsewhere. These questions would seem to indicate a possible need for an ability to deal with unforeseen circumstances, which will potentially affect a large number of services over a wide area, perhaps to the extent where an

individual or individuals will find it difficult to produce a reasonably optimal response, taking into consideration all or even most of the factors at play.

2.48 Later in his book, Hall (2001, p90) says that

*“ARS is likely to become much more widespread as train services become more regular and reliable. It produces a more predictable response and is capable of being programmed to make the optimum decisions in the event of interruptions to the service resulting from late running, cancellations and mishaps, etc.”*

It is not clear from this, however, that the necessary work has been done to enable ARS to make these optimal decisions. The same applies to other, competing systems available from such organisations as Union Switch & Signal (*Modern Railways*, July 2004, p21), for which similar claims are made, but whose operating details are commercially sensitive and confidential. Indeed, Pachl (2002, p208) concedes that “the development of systems to automatically solve schedule conflicts is still in a very early stage.” His view is supported by White (2003a, p. iv), who observes that “there have been attempts at automatic systems that will eliminate [rail traffic congestion]; however, none have been a great success.” In any case, a facility to manage disruptive incidents would be useful for planning railway operations around maintenance possessions, and for dealing with disruptive incidents in areas not covered by IECCs and ARS.

2.49 Furthermore, it would be very useful to identify from first principles (i) the key factors and variables for consideration in disruption management, (ii) which independent or decision variables are most susceptible to beneficial adjustment in different scenarios, and (iii) the various methods and strategies available and appropriate to make these adjustments in order to achieve the desired results.

2.50 In section 1.5 of its Capacity Utilisation Policy consultation document (2002), the Strategic Rail Authority (SRA) includes the following among its aims:

- *to find the best use that can be made of existing network capacity;*
- *to ... identify where enhancement investment in the network and its use is needed; and*
- *to ... determine the best use of any funds for capacity improvements to the network.*

Elsewhere (section 2.8), the document explicitly notes the benefits deriving from an “ability to achieve better service recovery following disruptive incidents.” This research project seeks to enhance that very ability, and to meet the first of the three aims above at times of disruption to normal operations. In the subsequent Network Utilisation Strategy document (Strategic Rail Authority, 2003, p12), the importance of “better train regulation ... to accommodate growing demand whilst also potentially improving performance” is noted. Suggested measures to improve regulation include the “revisit[ing of] prioritisation rules, class regulation practices and [the] use of passing facilities by passenger services.” The first two of these measures fall directly within the scope of this research project.

2.51 The principles identified and techniques developed during the course of the research are also likely to be of use in pursuing the second and third objectives listed above, in their potential use for simulating various disruptive scenarios, and the capability of proposed improvements to reduce their operational impacts.

### 3.0 THE DEVELOPMENT OF A SPREADSHEET MODEL FOR THE CALCULATION OF TRAIN JOURNEY TIMES

- 3.1 In order to determine the consequences of a disruptive incident for railway operations, it is necessary to evaluate the comparative effectiveness of alternative strategies for dealing with the incident. The delay experienced by trains is a fundamental measure of service quality, and delay minutes are used by Network Rail to measure performance and to target improvements (Network Rail, 2004b, p4). A delay minute is defined in the Office of Rail Regulation's online glossary (2004a) as follows:

*A measure equating to one train being delayed for one minute when compared with the timetabled journey time between two points.*

The delay at any point in the journey of a train may be calculated as the difference between the elapsed and the theoretical or scheduled journey times to that point. In order to be able to measure delay, it is therefore essential first to be able to calculate train journey times accurately.

- 3.2 The calculation of point-to-point train journey times is also required for many aspects of railway operations planning and management. These include: timetable compilation; evaluation of the impacts of changes to infrastructure and/or rolling stock standards; determination of rolling stock fleet sizes; and evaluation of the effects of the introduction of new services on the timing and reliability of existing ones. Martin (1999, p1287) confirms that "simulation provides a valuable tool in both the design of new infrastructure [and in] assisting in the process of translating a railways business aspiration into a technical specification." He observes that "growing sophistication with the available simulation systems is leading to the individual aspects of simulation being integrated to provide a comprehensive planning and development tool", but notes (p1288) that "despite all the added functionality [in commercial software packages,] the basis for calculation remains the run time between stations for each train type". The calculations involved are not particularly difficult, but they can be extensive, repetitive and laborious (and thus error-

prone) when performed manually, and so are very well suited to automatic computation.

- 3.3 This Chapter describes the updating and further development of an existing spreadsheet model for train journey time calculation. Following this introduction, the origins and specification of the model are described, including a list of the required improvements to its functionality. The development of the model to date is then described, including further explanations of the nature and purpose of the improvements introduced, and screenshots of the various input and output facilities. This is followed by a brief explanation of tractive effort, train resistance and line resistance, their effects on train performance, and the manner in which they have been used in the development of the model. The Chapter concludes with some potential avenues for further model development. A detailed description of the model structure and the algorithms employed are included in a separate volume as Technical Appendix A. This Appendix is confidential, and will not be included in the published EngD thesis.

### **Origins and Specification of the Model**

- 3.4 The Arup RUNTIME model was originally developed in 1993 as a SuperCalc spreadsheet macro, and was subsequently adapted, and updated and further developed, for use with the Quattro Pro spreadsheet application. Reflecting the increasing dominance of Microsoft's Office software, the first element of the author's EngD research activity entailed the development of an Excel-based version of the model, using the Visual Basic for Applications (VBA) programming language.

- 3.5 The original aim of the RUNTIME model was

*to provide a means of estimating station-to-station running times for passenger trains [by calculating] the time taken for a train to travel along a length of track given the train performance (acceleration, braking and maximum speed) and the distance travelled*



(Arup Transportation, 1993, pp1, 2). The total length of track along which a modelled train travelled (i.e. its route) was divided into segments of constant maximum line speed. Segment boundaries were additionally defined by any stops along a train's route; i.e. a single section of track with a constant line speed would be divided at an intermediate point into two route segments if the train was scheduled to stop at that intermediate point, which would typically represent a station. In addition to the train performance data (input in  $\text{m/s}^2$ ,  $\text{m/s}^2$  and  $\text{km/h}$ , respectively), the following information was input to the model for each segment of the route in question: segment length (km), start speed (km/h), end speed (km/h), maximum permitted line speed for segment (km/h) and, where applicable, the station dwell time (seconds) at the end of the segment.

3.6 These data were used with Newton's equations of motion to calculate the acceleration and deceleration times and distances, and the intermediate constant-speed travel ('cruise') time and distance, for each segment. If a segment maximum line speed exceeded the maximum train speed, the latter value was adopted as the segment maximum speed. In cases where the total required acceleration and deceleration distances exceeded the segment length, an iterative procedure was used to determine the maximum attainable speed, and the resulting acceleration and deceleration times and distances. Obviously, in such cases the cruise time and distance are both zero. In the case of segments ending at a station, a dwell time was added to the sum of the acceleration, cruise and deceleration times to obtain the total segment time. The individual segment times were then added together to determine total journey time. Enhancements made to the Quattro Pro version of the model included consideration of the effects of track gradient and curvature on train performance, and the provision of graphical output of

- (i) maximum vs. achieved segment speeds, and
- (ii) train graphs (i.e. plots of elapsed distance vs. elapsed time).

Extracts from the original Arup documentation, showing the macro structure and typical output from the model, are included in this document as Appendix A. (Note: the values of some of the output from the updated model displayed in the text differ from those shown in the Appendix; this is because some segment length and/or segment maximum speed values have been changed to test and/or illustrate various aspects of the updated model.)

- 3.7 Several areas for potential improvement were identified in the original documentation. These included facilities for consideration and analysis of the following: how trains should best be operated; traction energy consumption; alignment optimisation; train coasting; and the effects of gradients and curves. Other aims included the provision of on-line instructions and help facilities. Of these, the effects of gradients and curves were taken into partial consideration in the later, Quattro Pro-based model. Graphical outputs of maximum vs. achieved lines speeds and of distance travelled vs. time elapsed were also added to the Quattro Pro version.
- 3.8 The initial aim of this research activity was to reproduce the original features in the Excel version of the model, and then to introduce additional, specific improvements to the model, as identified and agreed in the course of discussions with Arup. These included:
- (i) Provision of a facility to specify segment lengths and train performance values in terms of various units, both Imperial (since much infrastructure data in Britain remain in Imperial units) and metric, thus avoiding the need for manual data conversion;
  - (ii) Provision of a facility to specify a ‘pathing time’ allowance (to compensate for possible delays at the approaches to junctions, for example) for any segment which requires it, reflecting industry practice in Britain;
  - (iii) Introduction of means to fully incorporate and calculate the effects of line curvature and gradient on maximum line speed and train performance;

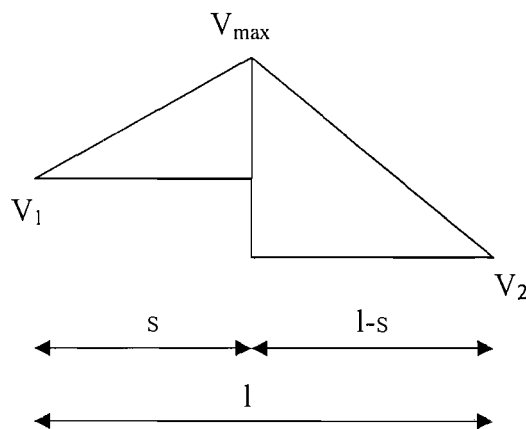
- (iv) Calculation and tabulation of the average speed for each route segment, and indication of the minimum average segment speed for a journey, highlighting the comparative performances of different segments of the route;
- (v) Provision of a facility to round segment travel times up to the nearest 30 seconds for output purposes, again reflecting industry practice in Britain (segment and total journey times are initially calculated within the model to the nearest second);
- (vi) Provision of the means of calculating directly the journey time between two intermediate stations or other specified points, without having to manually add or subtract segment travel times;
- (vii) The provision of continuous graphical output of Speed vs. Distance and vs. Time. This provides additional information about the journey in question, and particularly helps to highlight situations where train speed (and thus journey time) is constrained by maximum line speeds, and where train performance limitations prevent the full exploitation of available line speeds;
- (viii) Consideration of the ‘knock-on’ effects of a train not being able to achieve the specified end velocity for a segment, whereby the calculated segment travel time value would be misleading, and would invalidate the total journey time value. For example, if a train were unable to accelerate or decelerate sufficiently to achieve the specified speed at the end of a segment, the theoretical segment and overall travel times, based on the specified speed values, would be inaccurate, and would conceal a ‘speed discontinuity’ at the segment boundary, whereby the achievable speed at the end of one segment would not match the specified speed at the the start of the next. This is likely to happen only rarely in practice, but may occur on short segments where the train performance and/or infrastructure characteristics prevent the train from accelerating or decelerating sufficiently to achieve the specified end velocity value;
- (ix) Consideration of the constraining effects of train lengths on acceleration and speed as trains pass from segments with lower maximum line speeds

to ones with higher maximum values, where the acceleration of the train may be constrained by the lower line speed; and

- (x) The provision of facilities to determine and use non-linear rates of acceleration and deceleration, to specify the performance characteristics of locomotive-hauled passenger and freight trains of non-fixed formation, and to take account of trains ‘coasting’ (i.e. travelling without the application of power or brakes, restrained by the train’s inherent resistance to motion, by track curvature (if applicable), and by any track gradient acting against the progress of the train).

### Initial Development of the Excel-Based Model

3.9 The initial Excel version of the RUNTIME model was developed mainly by reference to the original 1993 model documentation. Although this documentation was partly superseded by subsequent model updates, it provides a useful account of the fundamental aims and principles of the model, and includes a worked example of the model in operation. This was helpful for validation purposes. The original model and the graphical output from the Quattro Pro version were readily reproduced in Excel, and the iterative approach to calculating maximum achieved segment speeds was replaced by a quicker direct calculation, using a formula derived as follows:



$$V_{max}^2 = V_1^2 + 2as \quad (\text{where } a = \text{train acceleration rate})$$

$$V_2^2 = V_{max}^2 - 2d(l - s) \quad (\text{where } d = \text{train deceleration rate})$$

Therefore,

$$V_1^2 + 2as = V_2^2 + 2dl - 2ds$$

$$2as + 2ds = V_2^2 - V_1^2 + 2dl$$

$$s = \frac{V_2^2 - V_1^2 + 2dl}{2(a + d)}$$

$$V_{\max}^2 = V_1^2 + 2a\left(\frac{V_2^2 - V_1^2 + 2dl}{2(a + d)}\right)$$

$$V_{\max} = \sqrt{V_1^2 + 2a\left(\frac{V_2^2 - V_1^2 + 2dl}{2(a + d)}\right)}$$

$$\text{Acceleration Time, } t_{acc} = \frac{V_{\max} - V_1}{a}$$

$$\text{Deceleration Time, } t_{dec} = \frac{V_{\max} - V_2}{d}$$

The formulae were tested and found to produce the correct results, irrespective of whether  $V_1$  is greater than, less than or equal to  $V_2$ . However, their use turned out to be short-lived, since the consideration of variable rates of acceleration and deceleration, as described below, required the development of an enhanced iterative procedure, capable of handling both linear and variable acceleration and deceleration rates.

- 3.10 This initial version of the model was used by Arup for the determination of fleet sizes for a proposed LRT system (now under development) for Edinburgh, based on calculated journey times and the specified service frequency, and for the assessment of possible improvements to heavy rail services between Leeds and Sheffield. Experience thus gained from the use of the model was fed back into its further development.

### **Subsequent Model Development**

- 3.11 Following the replication of the original model in Excel, work continued on the introduction of the identified improvements. Of these, the first six were

reasonably straightforward to introduce, numbers (vii) and (viii) were somewhat less so, and the last two were the most challenging.

### Input to the Model

- 3.12 A screenshot of the data input worksheet for the model is shown in Figure 3.1, below. The version shown accepts all the data required by the features (i) to (ix) above. The Input Table is generated by clicking the ‘Build Input Table’ button, having specified the required number of route segments. The train performance data are input to the spreadsheet cells above the four buttons. The model is run by clicking the ‘Calculate RunTime’ button, and the Input and Output Tables and their contents are cleared by clicking the lower left and right buttons, respectively.

Figure 3.1: RUNTIME Model ‘Input and Results’ Worksheet (Input Data)

The screenshot shows a Microsoft Excel spreadsheet titled 'Arup RUNTIME Model Temp.xls'. The worksheet is divided into two main sections: input parameters and a detailed 'INPUT DATA' table.

**Input Parameters:**

- Type of Train: 156 Sprinter
- Maximum Speed (km/h): 96.5
- Train Length (m): 46
- Acceleration Profile: Linear, Acceleration (m/s<sup>2</sup>): 0.5
- Deceleration Profile: Linear, Deceleration (m/s<sup>2</sup>): 0.5
- Total No. of Route Segments: 20
- Point-to-Point Output Data: Calculated

**Buttons:** Build Input Table, Calculate RunTime, Clear Input Table, Clear Results Table

**INPUT DATA Table:**

Segment No.	Segment Description	Length, S	Gradient	Ballast	Cant (mm)	Start Vel.	End Vel.	Max. Line Speed	Pathing Time (s)	Dwell Time at End (s)
	From	To	Unit: m	Unit: %	Unit: m	Unit: km/h	Unit: km/h	Unit: km/h		
1	Darlington		0			0	7.2	112.65408		0
2			18			7.2	18			0
3			27			18	27			0
4		North Road	36			27	36			0
5	North Road		45			36	45			0
6			54			45	54			0
7			63			54	63			0
8			72			63	72			0
9			81			72	81			0
10		Heistington	90			81	90			30
11	Heistington		100			90	100			0
12		Aycliffe	110			100	110			60
13	Aycliffe		120			110	120			0
14			130			120	130			0
15		Snickon	140			130	140			60
16	Snickon		150			140	150			0
17			160			150	160			0
18			170			160	170			0
19			180			170	180			0
20		B Auckland	190			180	190			0

- 3.13 Features (i) and (ii) required only minor changes to the model algorithms and VBA code. An example of the range of units available for data input is shown in column G of the worksheet shown in Figure 3.1; all data are subsequently converted to SI units for processing, with the exception of line speeds, which are

initially converted (if necessary) to km/h, in order to implement feature (iii). Pathing times are simply input to worksheet Column T, and added to the subsequently-calculated segment journey times. For the purpose of implementing feature (iii), the effects of track curvature and cant (equivalent to highway superelevation) on maximum line speed are determined by means of the appropriate equation in *British Railway Track* (The Permanent Way Institution, 1993, p372):

$$V_{\max} = 0.29(R(E_a + D_{\max}))^{0.5}$$

where

$V_{\max}$  = maximum train speed allowable for a given combination of R and  $E_a$  with  $D_{\max}$  (km/h);

R = radius of curvature of track (m);

$E_a$  = cant applied to the track (mm); and

$D_{\max}$  = maximum allowable deficiency of applied cant (mm).

R,  $E_a$  and  $D_{\max}$  are input to the table. If no value of R is entered, the relevant segment is assumed to be straight. If a radius is specified and the applied cant and/or allowable cant deficiency values are left blank, the latter values are assumed to be zero. The input/assumed values are then used to calculate the maximum allowable speed for the segment. If no maximum line speed value has been specified, the calculated value is adopted and displayed; if the calculated value is greater or less than the specified value, a dialog box is displayed, explaining the situation and giving the user the option of adopting the calculated value. Furthermore, if the specified track radius value is less than a 'sensible' (400m for heavy rail) or absolute (100m for heavy rail; 25m for light rail) minimum value, the user is warned accordingly, again by means of a dialog box, and is asked, or given the option, to change the specified value.

- 3.14 Input gradient values are used to calculate the component of a train's weight acting parallel to the track, which is then added to or subtracted from the train's linear acceleration and deceleration rates, according to the gradient 'direction'. Consideration of the effects of gradient on non-linear acceleration and

deceleration rates and on maximum train speeds is described later in this Chapter.

### **Model Output**

- 3.15 Screenshots of the summary tabular output from the model, including features (iv) and (v), are shown below in Figures 3.2(a) and 3.2(b). The contents of Figure 3.2(a) are described below, by worksheet column label.

Column W: an indication of whether the specified (or modified – see above) maximum line speed has been achieved. If the specified (or modified – see below) maximum train speed is less than the line speed, the user is also warned by means of a Dialog Box;

X and Y: the calculated individual and cumulative segment journey times, including any specified dwell times;

Z and AA: the values shown in Columns X and Y, plus any specified pathing allowances;

AB and AC: the values shown in Columns Z and AA, rounded up to the nearest 30 seconds, in accordance with UK practice;

AD: the cumulative distance to the end of each segment; and

AE and AF: the maximum and achieved line speeds. If the value for a segment in Column W is ‘Yes’, the two values will be equal; otherwise, the second column will show the maximum achieved speed for the segment.



Figure 3.2(a): RUNTIME Model Summary Tabular Output

Segment No.	Max. Line Speed Achieved?	Calculated Times (h:mm:ss) Segment	Cumulative	Incl. Path, All. (h:mm:ss) Segment	Cumulative	Rounded Times Segment	Cumulative	Cum. Dist (km)	Max. Line Speed (km/h) Limit	Achieved
1	No	00:08:17	00:08:17	00:08:17	00:08:17	00:08:30	00:08:30	12.000	112.7	96.5
2	Yes	00:00:30	00:08:47	00:00:30	00:08:47	00:00:30	00:09:00	12.140	18.0	18.0
3	Yes	00:00:16	00:09:03	00:00:16	00:09:03	00:00:30	00:09:30	12.230	27.0	27.0
4	Yes	00:00:11	00:09:14	00:00:11	00:09:14	00:00:30	00:10:00	12.320	36.0	36.0
5	No	00:00:09	00:09:23	00:00:09	00:09:23	00:00:30	00:10:30	12.410	45.0	43.2
6	Yes	00:00:59	00:10:22	00:00:59	00:10:22	00:01:00	00:11:30	13.275	54.0	54.0
7	Yes	00:01:55	00:12:16	00:01:55	00:12:16	00:02:00	00:13:30	15.495	72.0	72.0
8	Yes	00:00:55	00:13:11	00:00:55	00:13:11	00:01:00	00:14:30	16.535	72.0	72.0
9	Yes	00:00:17	00:13:28	00:00:17	00:13:28	00:00:30	00:15:00	16.877	72.0	72.0
10	Yes	00:02:41	00:16:10	00:02:41	00:16:10	00:03:00	00:18:00	19.107	72.0	72.0
11	Yes	00:00:31	00:16:41	00:00:31	00:16:41	00:01:00	00:19:00	19.348	54.0	54.0
12	Yes	00:02:52	00:19:33	00:02:52	00:19:33	00:03:00	00:22:00	21.148	72.0	72.0
13	Yes	00:01:10	00:20:43	00:01:10	00:20:43	00:01:30	00:23:30	22.148	72.0	72.0
14	Yes	00:01:42	00:22:25	00:01:42	00:22:25	00:02:00	00:25:30	24.148	72.0	72.0
15	Yes	00:01:28	00:23:53	00:01:28	00:23:53	00:01:30	00:27:00	24.348	48.0	48.0
16	Yes	00:00:58	00:24:52	00:00:58	00:24:52	00:01:00	00:28:00	24.948	48.0	48.0
17	Yes	00:00:12	00:25:04	00:00:12	00:25:04	00:00:30	00:28:30	25.109	48.0	48.0
18	Yes	00:01:23	00:26:26	00:01:23	00:26:26	00:01:30	00:30:00	26.208	48.0	48.0
19	Yes	00:02:25	00:28:51	00:02:25	00:28:51	00:02:30	00:32:30	28.908	72.0	72.0
20	Yes	00:00:22	00:29:13	00:00:22	00:29:13	00:00:30	00:33:00	29.028	32.0	32.0
		Calculated		Incl. Path, All.		Rounded				
		00:29:13 (h:mm:ss)		00:29:13 (h:mm:ss)		00:33:00 (h:mm:ss)				
		29.028 km		29.028 km		29.028 km				
		59.611 km/h		59.611 km/h		52.784 km/h				

Figure 3.2(b): RUNTIME Model Summary Tabular Output

SUMMARY RESULTS					
End Velocity (km/h)		Av. Segment Speed (km/h)			Notes
Specified	Achieved	Calc. Time	Incl. Path, All.	Rounded Time	
7.2	7.2	86.8	86.8	84.7	
18.0	18.0	16.9	16.9	16.9	
27.0	27.0	20.4	20.4	10.6	
36.0	36.0	29.0	29.0	10.8	
45.0	43.2	37.7	37.7	10.8	Specified End Velocity Not Achieved. Insufficient Acceleration and/or Segment Length. Accel
54.0	54.0	53.1	53.1	51.9	
48.0	48.0	69.3	69.3	66.3	
72.0	72.0	69.1	69.1	63.0	
72.0	72.0	72.0	72.0	41.0	
0.0	0.0	49.7	49.7	44.6	
54.0	54.0	27.9	27.9	14.5	
0.0	0.0	37.7	37.7	36.0	
72.0	72.0	51.4	51.4	40.0	
48.0	48.0	70.4	70.4	60.0	
0.0	0.0	8.2	8.2	8.0	
48.0	48.0	37.0	37.0	36.0	
48.0	48.0	48.0	48.0	19.2	
48.0	48.0	48.0	48.0	44.0	
32.0	32.0	67.2	67.2	64.8	
0.0	0.0	19.3	19.3	14.4	

3.16 The contents of Figure 3.2(b) are described below, again by worksheet column label.

AG and AH: the specified and achieved end velocities for each segment. As for the previous two columns, if the specified value for a segment is achieved, the values in the two columns will be equal. If, on the other hand, a train cannot accelerate or decelerate sufficiently (or maintain its speed) to achieve the specified value, as highlighted in the eighth item in the list of improvements to be made to the model, the user of the model requires (i) some notification, along with (ii) a means of resolving the situation, or, at least, some guidance as to how to do so. To achieve the first, a note to that effect is provided in Column AL (see below), indicating whether the train is unable to accelerate or decelerate sufficiently. The extent to which the specified end velocity value is under- or overshoot is indicated by the difference between the two values, enabling the start and end velocity values in the current and adjacent segments to be adjusted as necessary. Consideration was given to the automation of this process, so that start and end velocities would be automatically adjusted, but it was initially felt that it would be beneficial to leave such decisions in the hands of the user. Improvements to other aspects of the model led to a change in this respect, however, as is described later in this Chapter;

AI to AK: the average segment speeds, based on calculated times (including dwell times), calculated plus pathing times, and rounded calculated plus pathing times, respectively. The minimum average speed in each column is highlighted in bold. In this example, all three minimum values apply to the same segment, but this will not always be the case; and

AL: this column displays one (or none) of several standard messages, warning the user that, for example (and as in the case shown for segment 5), the specified end velocity cannot be achieved.

- 3.17 In addition to the various segment-by-segment outputs, the results are further summarised below the main table, as shown in Figure 3.2(a). These summaries show the total journey distance; the total calculated journey times, the total calculated journey times plus pathing allowances, and the total rounded values of calculated journey times plus pathing allowances, respectively; and the corresponding overall average journey speeds.
- 3.18 Another screenshot from the same worksheet is shown in Figure 3.3 on the following page, displaying the implementation of feature (vi), the calculation of intermediate journey times. This example shows the values for the calculated intermediate journey times, but the user is also given the option of displaying the calculated journey times plus pathing allowances, or the rounded overall values. This choice is made by selecting the appropriate option from the drop-down list indicated by 'Point-to-Point Output Data' on the 'Input and Results' worksheet, as shown in Figure 3.1.

**Figure 3.3: RUNTIME Model Intermediate Journey Times**

CALCULATED INTERMEDIATE JOURNEY TIMES (hh:mm:ss)																	
To End of Segment																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	00:08:17	00:08:47	00:09:03	00:09:14	00:09:23	00:10:22	00:12:18	00:13:28	00:16:10	00:16:41	00:18:33	00:20:43	00:22:25	00:23:53	00:24:52	00	
2		00:03:30	00:00:46	00:00:57	00:01:05	00:02:34	00:03:59	00:04:54	00:05:11	00:07:52	00:09:23	00:11:15	00:12:25	00:14:06	00:15:36	00:16:34	
3			00:09:16	00:00:21	00:00:26	00:01:34	00:03:29	00:04:24	00:04:41	00:07:22	00:07:53	00:10:45	00:11:55	00:12:39	00:15:05	00:16:04	
4				00:00:11	00:00:20	00:01:16	00:03:13	00:04:08	00:04:25	00:07:07	00:07:39	00:10:30	00:11:40	00:13:22	00:14:50	00:15:48	
5					00:00:09	00:01:07	00:03:02	00:03:57	00:04:14	00:06:55	00:07:26	00:10:18	00:11:28	00:13:11	00:14:39	00:15:37	
6						00:00:59	00:02:53	00:03:48	00:04:05	00:06:47	00:07:18	00:10:10	00:11:20	00:13:02	00:14:30	00:15:29	
7							00:01:55	00:02:49	00:03:07	00:05:49	00:06:19	00:09:11	00:10:21	00:12:03	00:13:32	00:14:30	
8								00:00:55	00:01:12	00:03:53	00:04:24	00:07:16	00:08:26	00:10:09	00:11:37	00:12:35	
9									00:00:17	00:02:58	00:03:30	00:07:32	00:09:14	00:10:42	00:11:41		
10										00:02:41	00:03:13	00:06:05	00:07:15	00:08:57	00:10:25	00:11:23	
11											00:00:31	00:03:23	00:04:33	00:06:15	00:07:44	00:09:42	
12												00:02:52	00:04:02	00:05:44	00:07:12	00:09:11	
13													00:01:10	00:02:52	00:04:21	00:05:19	
14														00:01:42	00:03:11	00:04:09	
15															00:01:28	00:02:27	
16																00:00:50	
17																	00
18																	
19																	
20																	

- 3.19 A considerable amount of intermediate data is generated during the preparation of the tabular output shown in Figures 3.2 and 3.3. These data are useful for

‘following through’ the calculation process and for ‘debugging’ changes to the model’s VBA code, and so are stored and displayed on the ‘Calculation’ worksheet (not shown). Further data are generated in order to produce the graphical output described below, including the second-by-second calculation of elapsed distances, ‘spot’ speeds and acceleration/deceleration rates, and segment maximum speed values that are used for the graphical display of train and line speed vs. distance and vs. time. These values are also stored and displayed on the ‘Calculation’ worksheet.

- 3.20 Figures 3.4 and 3.5 show screenshots of the updated versions of the graphical output provided in the Quattro Pro version of the model, while Figures 3.6 and 3.7 display the implementation of feature (vii) above. Figures 3.6 and 3.7 both show that the modelled train’s maximum speed is less than the maximum line speed on the first segment of the route, and Figure 3.7 clearly displays the linearity of the acceleration and deceleration rates. The graph shown in Figure 3.4 is used to highlight those route segments where trains are unable to achieve the maximum line speed, either because the train’s maximum speed is less than line speed, or because the segment length is insufficient for the train to accelerate to and/or decelerate from that speed. The data displayed in Figures 3.6 and 3.7 are obtained from a second-by-second calculation of elapsed time and distance, and ‘spot’ speed and acceleration, which are also used to provide the animation facility described later in this Chapter. Figure 3.8 shows an extract from the ‘Calculation’ worksheet containing these data. (Note: these second-by-second data are calculated on the basis of the calculated journey times only, exclusive of any pathing allowances and journey time rounding; this is because it would be impossible to allocate the additional times accurately within the segmental and overall journey times.)

Figure 3.4: 'Maximum and Achieved Line Speeds' Graph

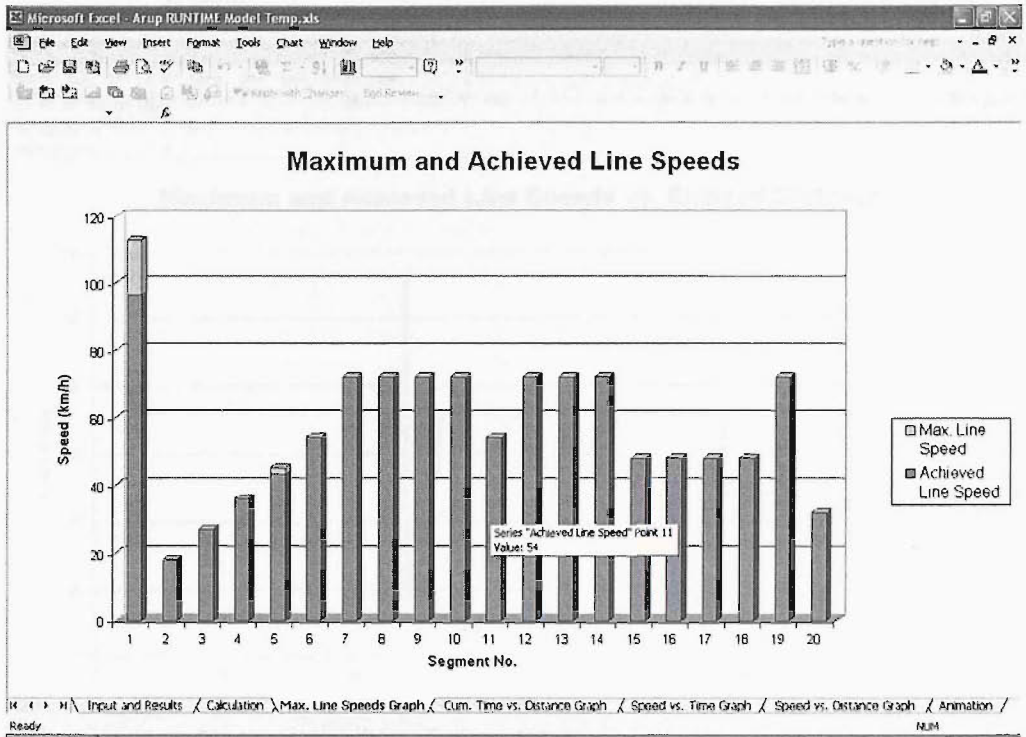


Figure 3.5: 'Cumulative Distance vs. Time' Graph

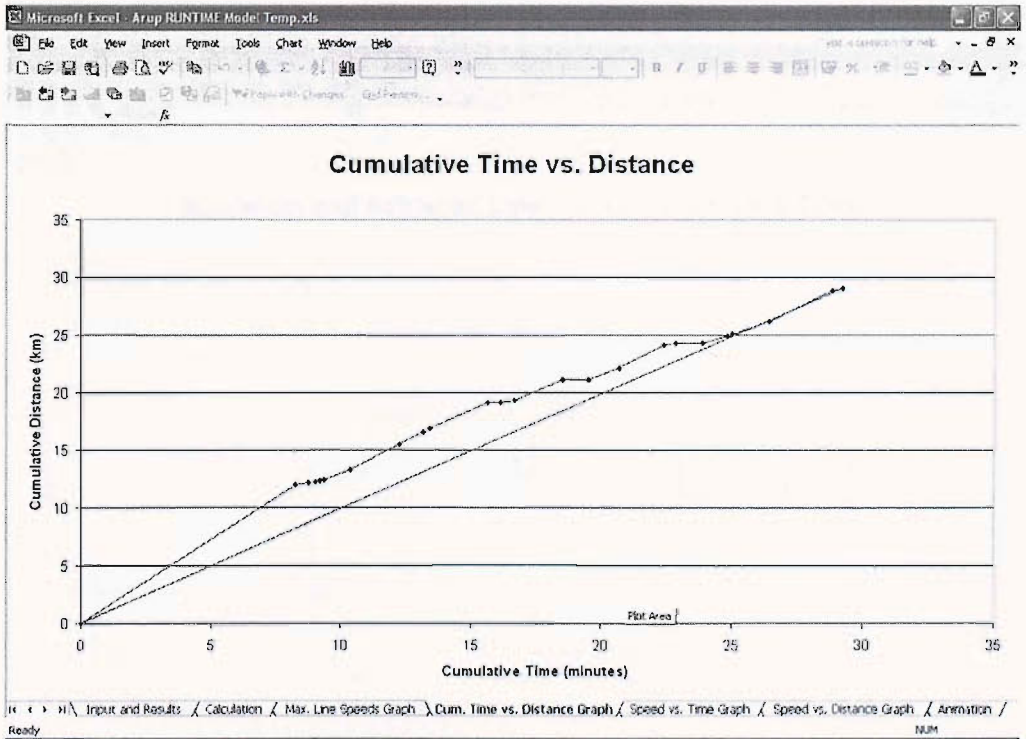


Figure 3.6: 'Maximum and Achieved Line Speeds vs. Elapsed Distance' Graph

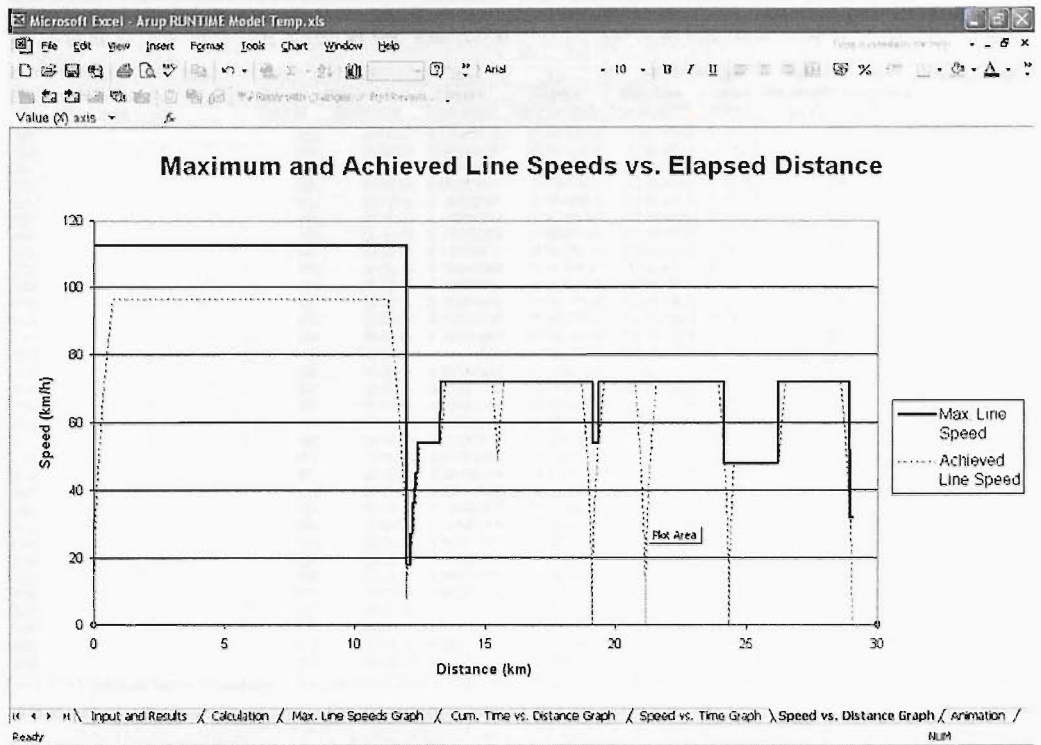


Figure 3.7: 'Maximum and Achieved Line Speeds vs. Elapsed Time' Graph

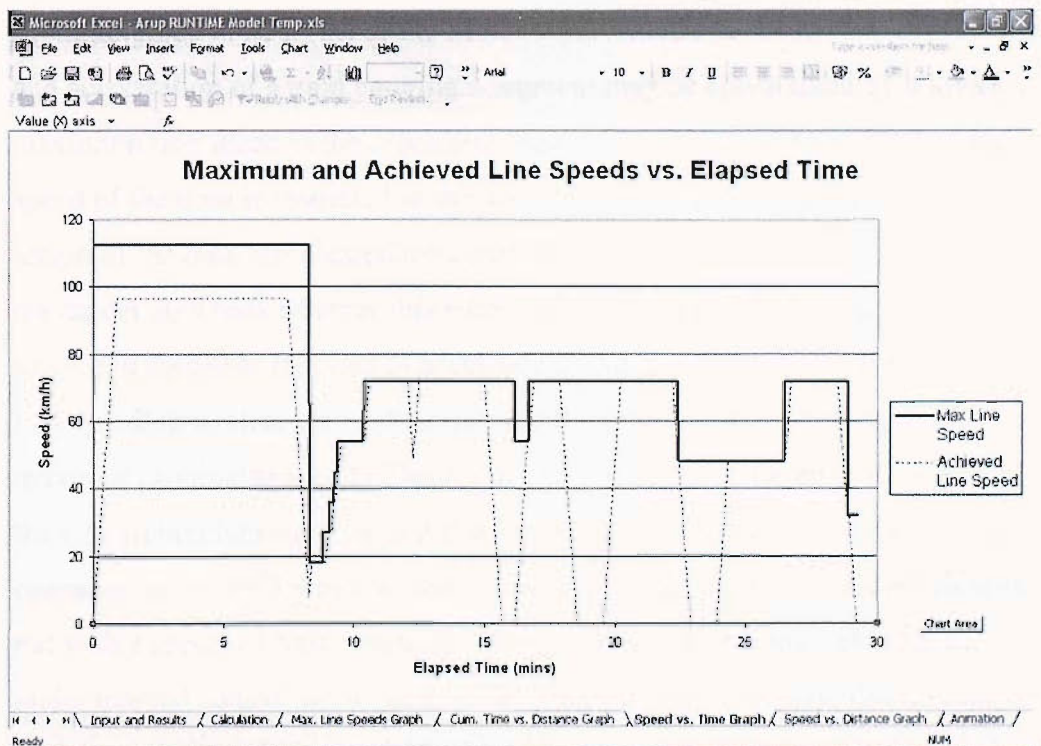


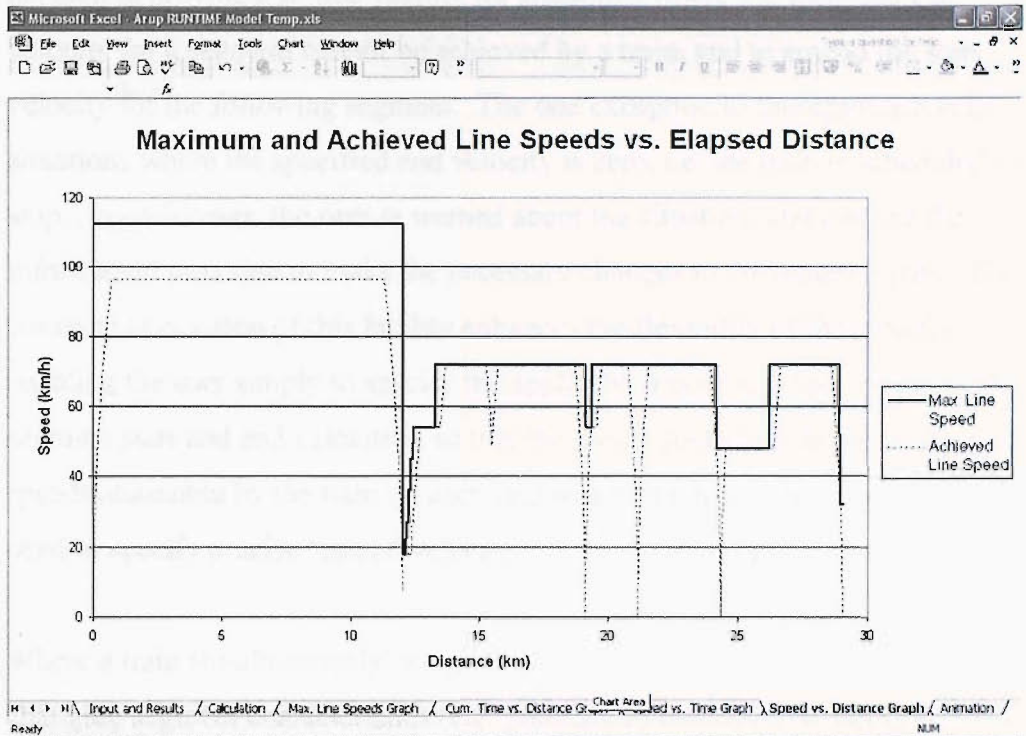
Figure 3.8: Extract from Second-by-Second Output Data

14	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BI
15	Cumulative Run Time (min)		Elapsed Time (s)	Elapsed Time (hh:mm:ss)	Elapsed Time (min)	Elapsed Distance (km)	Max. Line Speed (km/h)	Speed (km/h)	Acceleration (m/s <sup>2</sup> )	Segment No.		
499			482	00:08:02	8.033333778	11.9063158	112.6540833	34.97173	-0.5	1		
500			483	00:08:03	8.050000191	11.91902806	112.6540833	33.17173	-0.5	1		
501			484	00:08:04	8.066666603	11.92602053	112.6540833	31.37173	-0.5	1		
502			485	00:08:05	8.083333015	11.93652501	112.6540833	29.57173	-0.5	1		
503			486	00:08:06	8.100000381	11.94448949	112.6540833	27.77173	-0.5	1		
504			487	00:08:07	8.116666794	11.95195397	112.6540833	25.97173	-0.5	1		
505			488	00:08:08	8.133333206	11.95891844	112.6540833	24.17173	-0.5	1		
506			489	00:08:09	8.149999619	11.96538194	112.6540833	22.37173	-0.5	1		
507			490	00:08:10	8.166666985	11.97134642	112.6540833	20.57173	-0.5	1		
508			491	00:08:11	8.183333397	11.9768109	112.6540833	18.77173	-0.5	1		
509			492	00:08:12	8.199999809	11.98177538	112.6540833	16.97173	-0.5	1		
510			493	00:08:13	8.216666222	11.98623986	112.6540833	15.17173	-0.5	1		
511			494	00:08:14	8.233333588	11.99020433	112.6540833	13.37173	-0.5	1		
512			495	00:08:15	8.25	11.99366881	112.6540833	11.57173	-0.5	1		
513			496	00:08:16	8.268666412	11.99663329	112.6540833	9.771728	-0.5	1		
514			497	00:08:17	8.283333778	11.99909776	112.6540833	7.971728	-0.5	1		
515			498	00:08:18	8.300000191	12.00122416	18	8.228265	0.5	2		
516			499	00:08:19	8.316666603	12.00375979	18	10.02827	0.5	2		
517			500	00:08:20	8.333333015	12.00679542	18	11.82827	0.5	2		
518			501	00:08:21	8.350000381	12.01033105	18	13.62827	0.5	2		
519			502	00:08:22	8.366666794	12.01436688	18	15.42827	0.5	2		
520			503	00:08:23	8.383333206	12.01890231	18	17.22827	0.5	2		
521			504	00:08:24	8.399999619	12.02395645	18	18	0	2		
522			505	00:08:25	8.416666985	12.02888645	18	18	0	2		
523			506	00:08:26	8.433333397	12.03385645	18	18	0	2		
524			507	00:08:27	8.449999809	12.03885645	18	18	0	2		
525			508	00:08:28	8.466666222	12.04385645	18	18	0	2		
526			509	00:08:29	8.483333588	12.04885645	18	18	0	2		
527			510	00:08:30	8.5	12.05385645	18	18	0	2		
528			511	00:08:31	8.516666412	12.05885645	18	18	0	2		
529			512	00:08:32	8.533333778	12.06385645	18	18	0	2		
530			513	00:08:33	8.550000191	12.06885645	18	18	0	2		

### Segment Boundary Issues

3.21 As highlighted in item (ix) of the listed improvements to the model, the speed and acceleration of a train entering a segment may be constrained by a lower maximum line speed in the preceding segment(s). In such cases, the maximum speed of the train is restricted to the lowest applicable line speed until the entire length of the train has cleared the constraining segment(s). For each segment, the model now tests whether this situation applies, and, if so, calculates the times and distances required to accelerate to the preceding segment maximum line speed(s), to clear the segment(s), and then to accelerate to the next maximum attainable speed. These values are then added together to determine the overall acceleration time and distance values. This feature can be seen in operation in Figure 3.9 below, based on the same data as the preceding figures, but with a specified train length of 200m. In practice, it is unlikely that a train under manual control would accelerate to such a precise pattern; discussions with operators would be useful to determine any variations in typical operating practice, which could then be included as options within the model for determining journey times under different operating practices.

**Figure 3.9: The Constraint of Acceleration by Preceding Segment Speed Limits**



- 3.22 As noted above, consideration was given during the development of the model to providing an automatic response to situations where the specified end velocity for a segment could not be achieved, so that the specified values would be replaced by appropriate alternatives. Although a user of the model would have been informed that this was happening, and why, it was initially considered that it would be preferable to allow users to **make the necessary** changes manually, having been given the necessary diagnostic information, to ensure that the reasons for the changes were fully understood and ‘thought through’.
- 3.23 However, the implementation of the feature described in para. 3.21 means that the starting velocity for a segment may be constrained by the maximum line speed for any preceding segment within the length of the specified train type (apart from the immediately preceding one, for which any discrepancy in specified speeds would be picked up and corrected during the initial input data validation checks). If/when this happens, the user is informed and given the option of adopting the amended values calculated by the model, in order to



produce a valid set of results. In the interests of consistency, it was therefore decided to provide a similar facility for situations where the specified end velocity for a segment cannot be achieved by a train, and to amend the start velocity for the following segment. The one exception to this approach is in situations where the specified end velocity is zero, i.e. the train is scheduled to stop. In such cases, the user is warned about the situation, and can use the information provided to make the necessary changes to the model inputs. The consistent provision of this facility enhances the flexibility of the model by enabling the user simply to specify the applicable maximum line speeds as the segment start and end velocities, so that the model seeks to achieve the highest speeds attainable by the train on each segment of the route, thus avoiding the need to specify precise ‘target’ velocities at the segment boundaries.

- 3.24 Where a train simultaneously occupies multiple segments, the effects of changing segment characteristics (e.g. gradient, as noted above, but also curvature, as described below) on train performance are also an issue, as the train traverses the various segments (see also para. 3.37).

### **Variable Rates of Acceleration and Deceleration**

- 3.25 The most fundamental aspect of the further development of the model entailed the consideration of variable rates of acceleration and deceleration (item (x) in the list of improvements above), in addition to the linear rates previously employed. The use of constant average linear rates of acceleration is appropriate for some applications, such as Light Rail, and can also provide useful ‘first pass’ journey time indications in other situations, but the use of a non-linear rate will usually be more accurate. Trains in Britain (and the rest of Europe) typically brake at a maximum constant service rate of 0.09g (Tunley, 2003, p1463), but a variable rate of deceleration is required for detailed consideration of coasting behaviour. The key factors in any treatment of these issues are Tractive Effort and Train Resistance, and the effects of track gradient and curvature (grade resistance and curve resistance, respectively).

- 3.26 Tractive Effort is the force available to accelerate a train, and is a function of power and speed:

$$\text{Tractive Effort} = \text{Power/Speed}$$

However, the maximum value of tractive effort is limited by the weight on a locomotive's (or train's) driven axles and the coefficient of adhesion between wheel and rail, which typically has a value of about 0.2.

- 3.27 Train Resistance, as the term suggests, is the measure of a train's resistance to motion; it varies with the characteristics of a train and with speed, is also measured in units of force and is calculated by means of the Davis equation:

$$R = a + bV + cV^2$$

where

R = Train Resistance;

V = Train Speed; and

a, b, and c are constants determined by the train characteristics, obtained from empirically-derived formulae or directly from rolling stock tests.

- 3.28 The first two terms of the equation relate respectively to 'stiction' and rolling resistance, and depend on train weight and design and (for the second term) on speed. The third term depends on rolling stock size and shape and on the square of a train's speed, and relates to wind resistance. On trains of varying composition, overall resistance can be determined by summing the individual resistances of the locomotive(s) and trailing vehicles (Morlok, 1978, pp121, 122).

- 3.29 Total resistance is determined by adding the effects of line resistance, chiefly comprising the effects of track gradient and curvature, to train resistance. As noted above, grade resistance is the component of train weight acting parallel to the vertical alignment of the track, while curve resistance is due to various wheel-rail frictional forces and to the fact that railway vehicle axles are solid

and wheels slip slightly on both rails on curves. The formulae employed for curve resistance take different forms, two of which are shown below:

(i):  $r_{\text{curve}} = 650/(R-55)$  with  $R > 300\text{m}$

$$r_{\text{curve}} = 500/(R-55) \text{ with } R < 300\text{m}$$

where

$$r_{\text{curve}} = \text{specific curve resistance (\%o or N/kN)}$$

$$R = \text{radius (m)}$$

(Pachl, 2002, p31)

(ii)  $\text{Curve Resistance} = 0.8\text{lb/ton}^\circ$

where

$$^\circ = \text{angle subtended by a 100ft chord of curve}$$

(Hay, 1982, p146)

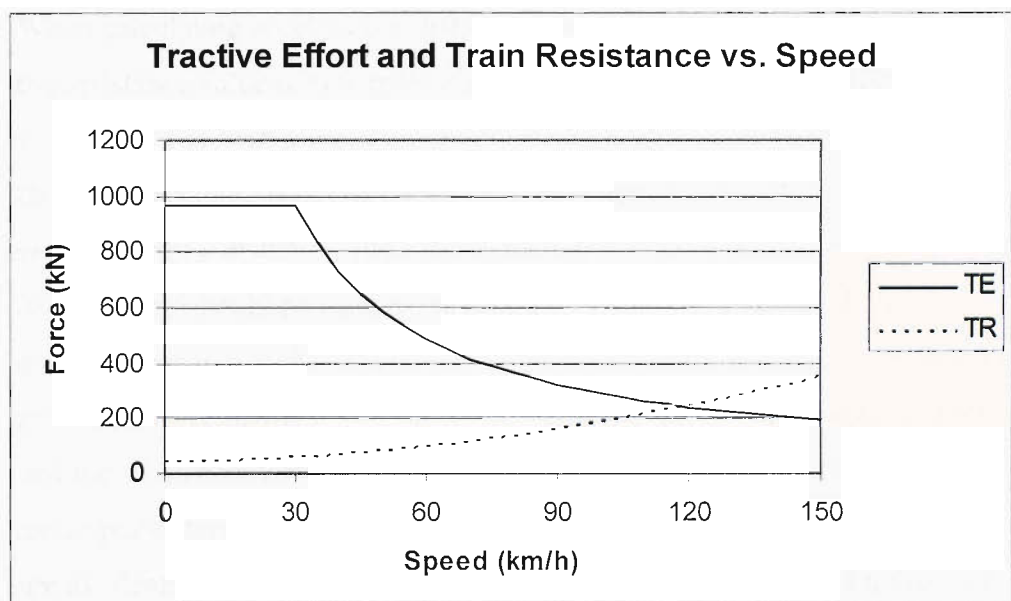
Comparison of the values obtained from the two formulae indicates close agreement for a wide range of curve radii and train weights. The formulation in (i) is a simplification for standard (1.435m) gauge applications of the more general Von Röckl formula (Dalla Chiara and Gačanin, 2004, p24). Other formulae include those proposed by Skramm and Desdonits (ibid).

3.30 It was noted earlier in this Chapter that different curvature restrictions apply to Heavy and Light Rail track alignments, 100m and 25m respectively being the absolute minimum radii that can be employed. It can be seen, however, that the formulae in (i) above yield nonsensical results for radii less than 55m, and that a radius value of 55m would result in division by zero, causing the model to ‘crash’, at worst, or, at best, the run to be terminated. It is also likely that radius values greater than but close to 55m would produce unreliable results. Since both formulae explicitly apply to Heavy Rail applications in any case, they may

not be suitable for Light Rail use, and searches and queries yielded no Light Rail equivalent. Following discussions with Arup, it was decided that it would be reasonable to assume that constant average linear rates of acceleration and deceleration apply in all Light Rail situations, and that the effects of curve resistance could be ignored in these cases. The model has therefore been set up in such a way that radii of less than 100m cannot be employed in conjunction with non-linear rates of acceleration or deceleration, and any attempt to do so results in the model run being terminated, and the user being advised by means of a dialog box as to why this is happening.

- 3.31 Typical plots of Tractive Effort and Train Resistance are shown in Figure 3.10 below. These are for a train with a total weight of 3814 tons, hauled by four 3300hp diesel-electric locomotives (Morlok, p156), with conversion to metric units. The flat portion of the Tractive Effort curve represents the limiting value of the weight on the locomotives' driven axles multiplied by the coefficient of adhesion, as described above, while the curved portion represents the division of the power available at the rail/drawbar (i.e. theoretical power minus transmission losses, etc.) by train speed. In practice, more detailed tractive effort data may be specified by locomotive/train manufacturers. The resistance curve represents train resistance as calculated by the Davis equation, again as described above.

**Figure 3.10: Tractive Effort and Train Resistance vs. Speed**

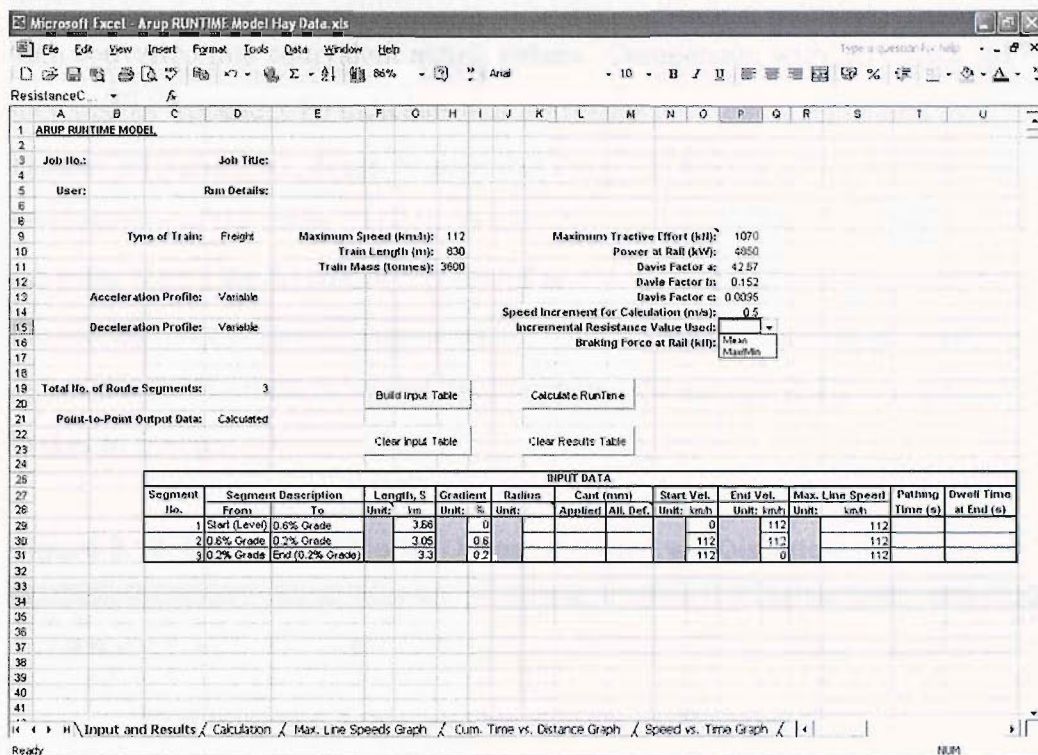


At any given speed, the net force available to accelerate the train is equal to the difference between tractive effort and total resistance. The maximum speed the train can achieve is that at which tractive effort and total resistance are equal, approximately 120 km/h in the example shown; this is known as the ‘Balancing Speed’.

- 3.32 Using the foregoing formulae, tractive effort, total resistance and net accelerating force can be calculated for any speed value (note: in certain circumstances, particularly on uphill grades, the net accelerating force may be negative, such that a train is forced to decelerate to a lower balancing speed; this eventuality is catered for in the model). Using, say, 1 km/h speed increments, incremental acceleration times and distances can then be calculated using Newton’s equations of motion and the average net accelerating force for each increment; the total acceleration time and distance values are then obtained by summing the incremental values. This is an approximate method, but, according to Pachl (p35), “it is impossible to calculate the speed curve of a train in an analytic manner [and] the speed curve can only be approximated step by step in [the] form of a sequence of straight line portions.” A similar approach can be used to determine braking times and distances (omitting tractive effort and adding a braking force to the total resistance), and coasting times and distances (by omitting both tractive effort and the braking force).
- 3.33 When calculating acceleration distances and times, Hay (p160) suggests using the resistance value for the maximum rather than the average increment speed, to produce a conservative result, but both Pachl (p36) and Morlok (p156) use the average value. Hay does not include an explicit calculation of braking and/or coasting distances and times, and makes no suggestion as to what value of resistance should be used in these cases. To be consistent with his conservative approach to the acceleration calculations, it would seem that the minimum incremental resistance value should be used for braking calculations, and the maximum value should be used for coasting calculations. These assumptions have been incorporated in the RUNTIME model, and both options are available to users, as can be seen in the screenshot shown in Figure 3.11: adoption of the ‘Max/Min’ option for the ‘Incremental Resistance Value Used’

input results in the use of conservative resistance values, while the ‘Mean’ option results in average values being used throughout. This Figure also shows the changes to the data input requirements when non-linear rates of acceleration and deceleration are used (the value of braking force used in this example is obscured by the drop-down list; a value of 3274 kN was employed).

**Figure 3.11: Data Input for Variable Acceleration and Deceleration Rates**



3.34 If a linear deceleration rate is specified in conjunction with a variable rate of acceleration, the braking force value is omitted, and the deceleration rate is input directly, as in Figure 3.1. If a linear acceleration rate is employed in conjunction with a variable rate of deceleration, however, all the performance data (power, tractive effort and Davis equation coefficients) are still required in order to calculate the balancing speed and check it against the specified line and train speed parameters.

## Model Validation

3.35 Attempts were made by the author and by Arup to obtain suitable data from the British rail industry with which to validate the model. These attempts were unsuccessful, however, so data from Hay (pp161-177) were input to the model, as shown in Figure 3.11. Figures 3.12 and 3.13, below, show some of the output obtained. The train performance parameters used by Hay are not explicitly stated in his book, and so were estimated on the basis of the information provided, and then converted into equivalent metric values. Comparison with his output (p177; included as Appendix B) indicates that his results are closely replicated by Figures 3.12 and 3.13. It can be seen that the train initially accelerates on the level section of the route, but does not reach its balancing speed for that part of the route (105.9 km/h) before it is forced to decelerate by the 0.6% gradient on the second route segment. Acceleration is again possible once the gradient is reduced to 0.2% on the third and final segment of the route, before the train brakes to a stop.

Figure 3.12: Example of Model Output – Speed vs. Distance

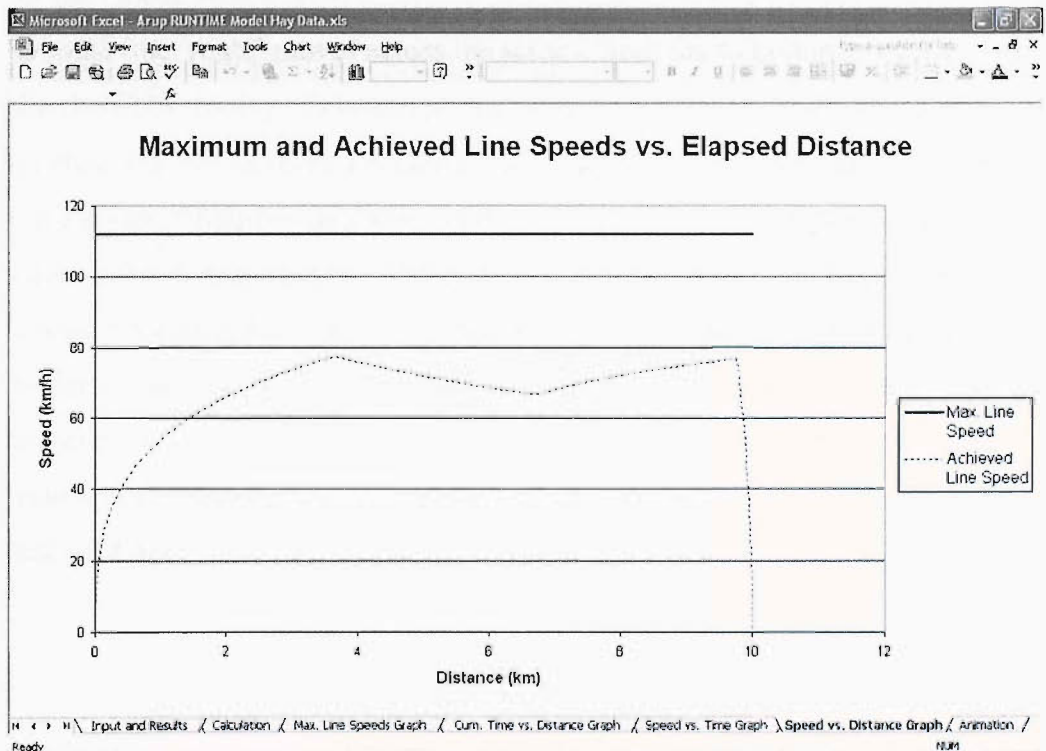
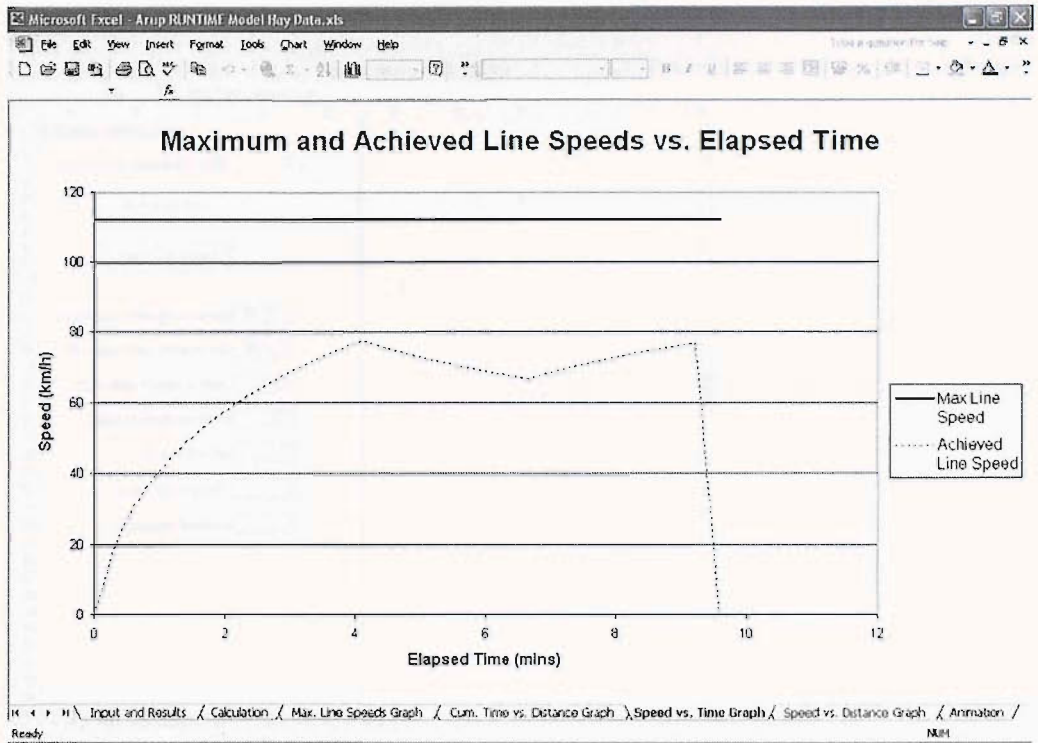


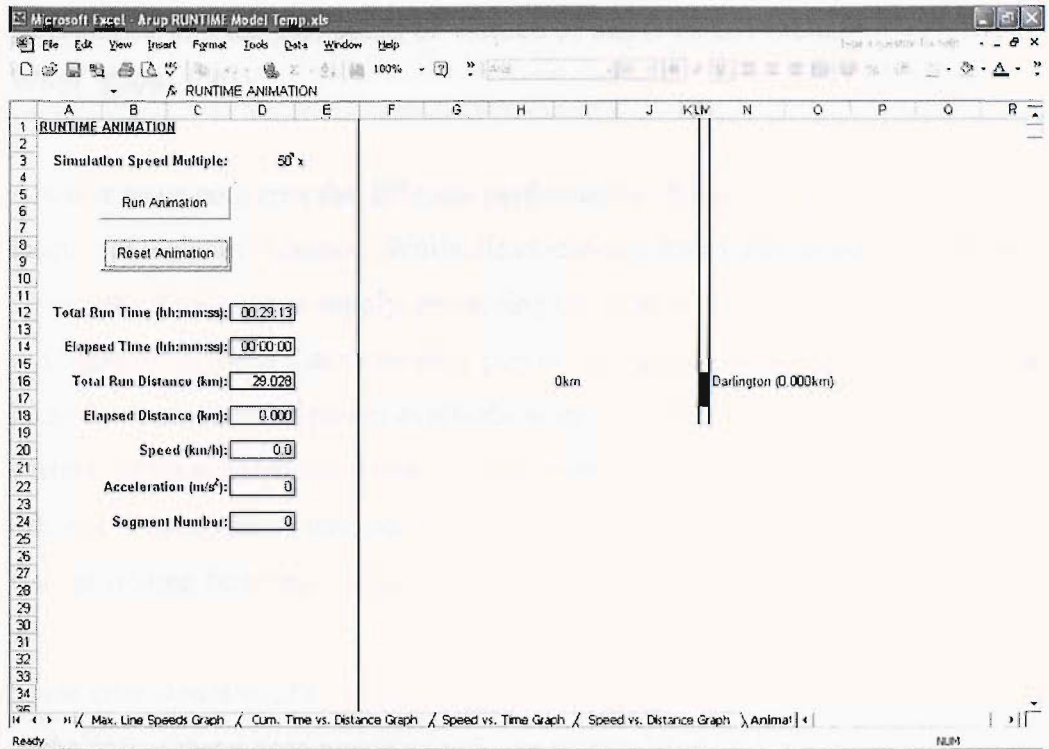
Figure 3.13: Example of Model Output – Speed vs. Time



3.36 In addition to the facilities described above, an animation facility was developed for the model, which provides a graphical display of the ‘train’ progressing along its route (the ‘route’ moves across the screen, from top to bottom), with a simultaneous display of elapsed journey time and distance, ‘spot’ speed and acceleration (which takes a negative value during deceleration), and the current route segment number, as shown below in Figure 3.14. The figure shows the information displayed at the start of a ‘run’; it was not possible to capture a screenshot during the animation process. Apart from an illustration of the modelled route, the facility provides little additional information to the user; however, it was a useful exercise for the author in the development of simulation techniques. The set of data used for the purposes of the animation is the same as that used in Figures 3.6 and 3.7, or 3.12 and 3.13.



**Figure 3.14: RUNTIME Animation Facility**



## Future Model Development

- 3.37 The model improvements specified by Arup have now been implemented, apart from the ability to specify train performance data in a range of different units, but there remain some other aspects of the model which could potentially be improved. Further feedback from users may result in changes to the model structure, and particularly to the feedback and guidance when input or other errors occur. Another outstanding issue is the consideration of the effects of variations in track gradient and curvature on trains as they cross segment boundaries, as noted earlier. It should be possible to approximate these effects, by taking average values of gradient and curvature, in conjunction with the calculations used for the consideration of the constraining effects on train performance of preceding segment maximum speeds.
- 3.38 There is also a need to consider the issue of Continuous Tractive Effort, i.e. the maximum value of tractive effort, corresponding to a minimum speed value, that can be sustained without overheating of traction motors. If the balancing speed for a route segment is less than the speed corresponding to the continuous

tractive effort rating for the train or locomotive(s), additional motive power may be required, and the user could be warned of this (Fox & Pritchard, 2004, p9; Woof, 1999, 2001).

- 3.39 Another issue concerns the different performance characteristics of diesel-electric and electric traction. While diesel-electric trains and locomotives have an essentially fixed power supply, producing the type of tractive effort curve shown in Figure 3.10, trains that draw their power from external electricity supplies can increase the amount of power available to them (within the limits of the supply system) as their speed increases, so as to maintain a constant value of tractive effort to higher speeds than are typically possible for diesel-electric equipment, thus providing faster acceleration (Fox & Pritchard, 2004, p9).
- 3.40 Some consideration of coasting behaviour is possible with the model as it stands, to the extent that a zero braking force can be specified, and the time and distance required to coast from the maximum segment speed to the specified end velocity can be determined. However, it would be useful, and probably more realistic, to be able to specify an interval of coasting between acceleration/cruising and braking, particularly for the purposes of optimising power consumption and making the best use of available pathing time allowances, where these remain available to a train's driver on the approach to the end of a route segment.
- 3.41 Beyond these issues, Arup's longer-term aspirations for the model were listed earlier, some of which have already been addressed in the course of the work described above; the outstanding longer-term objectives include facilities for the consideration and analysis of the following:
- (i) how trains should best be operated;
  - (ii) traction energy consumption;
  - (iii) alignment optimisation; and
  - (iv) train delay assessment for different operating scenarios and track layouts.

Other aims include the provision of further improvements to on-line instructions and help facilities.

## **4.0 THE DEVELOPMENT OF A GENERAL MODEL FOR THE SIMULATION OF RAILWAY OPERATIONS**

- 4.1 In order to achieve a realistic simulation of the operation of multiple trains in a railway network, a more general model than the one described in the previous Chapter is required. Such a model facilitates the simulation and the examination of the interactions between trains using the network, and also the interactions between the trains and the network's signalling system. There are many such commercial models available on the market (e.g. VISION, RailSys, OpenTrack, TrainPlan), at least some of which claim to be able to simulate disruptive incidents and responses to them (it is also claimed, as noted in Chapter 2, that ARS and a control system supplied by Union Switch & Signal have disruption management facilities). However, a user of these models is almost certainly restricted to using the facilities and algorithms provided by a model's developer, and will not have access to the workings of those algorithms or the flexibility to introduce and investigate alternative approaches.
- 4.2 In addition to their usefulness for developing and assessing responses to simulated disruptive incidents, such models have a much wider applicability to railway operations planning and are widely used by the railway industry and consultants; Arup already makes use of VISION. The commercial models generally offer a comprehensive range of simulation facilities, but are expensive to buy and maintain, and require considerable training and experience for staff to gain familiarity and comfort with them. For smaller pieces of work, particularly where it is of a 'first pass' or preliminary nature, simpler, less comprehensive and cheaper models have a role. This Chapter describes the development of a general model for the simulation of railway operations which is intended to satisfy the needs of the overall EngD research topic, and also some of Arup's simulation requirements.
- 4.3 Following this introduction, the origins and specification of the model are described. The approaches adopted to the description and definition of the modelled network, the signalling system and the timetable are then described. The application of the specified timetable to the specified network is then examined, including the handling of the interactions between trains and signals,

and the calculation of delay. Finally, some options for the further improvement of the model are considered. A detailed description of the model structure and the algorithms employed are included in a separate volume as Technical Appendix B. Again, this Appendix is for examination purposes only, and will not be included in the published EngD thesis.

### **Origins and Specification of the Model**

- 4.4 In order to meet the original objective of this research project, a model is required to simulate railway operations under perturbed conditions, and to investigate and assess the relative effectiveness of different responses. Such a model could take many forms, from a fully-animated representation of the simulated conditions to a non-graphical ‘black box’ representation, based on segmental running times and the re-scheduling of trains to reduce and/or minimise the overall delay.
- 4.5 In the event, the form of the model developed reflects Arup’s aims and aspirations in this area; i.e. to have an in-house model capable of easily and cheaply simulating operations in comparatively small sections of a railway network. Discussions with interested parties in Arup produced the following specification:
- Network and timetable data should be input to the model as files, rather than being ‘hard coded’ into the model, as was the case with an earlier, preliminary version.
  - Simulation results should be output to file(s).
  - The model should provide a visual representation of the network, showing train movements and signal aspects. The network should be represented as a set of straight lines: curves should not be represented graphically. Because of its potential usefulness, this graphical output should be a primary consideration, rather than a ‘bolted-on’ afterthought.
  - The model should initially be capable of simulating conventional 3- and 4-aspect colour light signalling, but should have the potential eventually to

simulate signalling systems which dispense with conventional lineside equipment (e.g. Levels 2 and 3 of ETCS (the European Train Control System)).

- The model should be capable of calculating the delay experienced by trains as a result of disruption and/or poor timetabling.

These are the main elements of the initial specification; other desirable features, which may be implemented in the longer term, include:

- The generation of train graphs as part of the model output (similar to the graph shown in Figure 3.5).
- A facility to ‘draw’ networks on the screen (direct manipulation of the input), rather than using extensive numerical input.
- Full area simulation for timetable planning.
- Analysis of power requirements, supply and distribution for electrified rail networks. Ideally, the power supply and distribution requirements should be assessed and determined directly, on the basis of a specified network and service pattern, rather than on a ‘trial and error’ basis, whereby different power supply and timetable options are tested in order to see which of the options yields the best results.
- Generation of rolling stock and train crew diagrams.

As mentioned in the first bullet point, a preliminary model was developed in the early stages of the research project, using the Visual Basic 6 programming language. While this model provided some useful initial experience in the area, its form was inflexible and it had only limited scope for improvement. It was decided, therefore, to start afresh with the object-oriented programming language Visual Basic .NET. An object-oriented approach is particularly suited to the needs of the model in question, and the data and simulation structures used provide a robust basis for future improvement and expansion of the model.

## Track Network Description and Definition

- 4.6 A fundamental requirement of the model is a facility to describe and depict the infrastructure comprising the relevant section of a railway network in an accurate and efficient manner. The infrastructure components which are of most interest and relevance are the track and the signalling system. Bridges, tunnels, etc. have no direct affect on the simulation process at the level of detail being considered, although it is desirable to have a facility to indicate their locations, so as to distinguish visually between grade-separated and flat junctions, for example. The obvious approach to simulating a network is to represent it by nodes and links; consideration was initially given to representing signals as nodes, but it was decided that such an approach would be inflexible, particularly for the consideration of different signalling options on a fixed track layout. Track and signals are therefore defined separately, in individual files, and the signalling system is ‘overlaid’ on the track network. The approach adopted for the definition of the signalling system is described in the next section of this Chapter, and the definition of the track layout is covered in the following paragraphs. Visual depiction of station names, layouts (platforms, etc.) and platform numbers is desirable, and should be reasonably straightforward, but has not yet been implemented.
- 4.7 Having decided to separate the definitions of the track layout and the signalling system, a further decision had to be made with respect to representing the track layout. There were two major options: (i) a link-based representation, with each link in the network being explicitly defined and described, in terms of its position and length; and (ii) a node-based representation, with the position, type and connectivity details of each node in the network being described, and the connectivity details enabling the specification of the links comprising the network. The node-based option was chosen for several reasons: it enables a more compact and flexible representation of the necessary data; timetable data is inherently node-based, and can thus readily be ‘mapped on’ to a node-based network; and navigation through a network can easily be defined in terms of a node-to-node progression, as defined by the specified timetable data.

4.8 The data structure adopted for the purposes of defining the track network is summarised in Table 4.1.

**Table 4.1: Data Used to Define Track Network**

Data Item	Comment
Total Number of Nodes in the Network	Not strictly necessary, but can be used to check for consistency of the subsequent data.
For each Node in the Network:	
Node Number	Unique identifying number.
X-Coordinate	Horizontal position (pixels) of node on screen.
Y-Coordinate	Vertical position (pixels) of node on screen.
Type	Junction, Terminus, Entry/Exit, etc.
Number of Linked Nodes	The number of nodes to which the current node is linked; i.e. the number of links from it.
List of Linked Node Numbers	Definition of the links from the node, and thus its role in the 'connectivity' of the network.
For each Linked Node:	
Link Label	A label describing the link between the two nodes (e.g. 'Up Main', 'Down Branch').
Start Chainage	The chainage, or specified distance along a route, of the 'Current Node' end of the link.
End Chainage	The chainage, or specified distance along a route, of the Linked Node end of the link.

As noted in Table 4.1, it is not strictly necessary to specify the total number of nodes in the modelled network, as this can be determined by counting the total number of listed nodes, but it does enable the comparison to be made between the specified and the counted total values, thus providing a consistency check. Such a check has not been implemented in the model as it stands, but the specified total is used for other purposes within the model. The screen coordinates have their origin (0, 0) in the top left corner of the screen, in contrast to easting and northing values, whose origin is at or beyond the 'bottom left' of the area under consideration. Any specified easting and northing (or other coordinate) values therefore need to be converted into suitable pixel values for display on the screen. This has to be done manually (or externally, anyway) in the current version of the model, but an obvious improvement would be the automatic conversion of specified coordinate values into values and a scale



suitable for display. The node type data item is not widely used in the current version of the model, but has a range of potential future uses, as described in the Further Work section of this Chapter.

- 4.9 As in the case of the total number of nodes comprising the network, it is not strictly necessary to specify the number of nodes linked to the ‘current’ node, since this can be determined by counting the number of specified linked nodes. Again, however, the specified value is used for other purposes within the model, and provides the possibility of introducing a check for data consistency. At the model’s current state of development, links can be defined in either direction, i.e. it makes no difference which of the two nodes at the ends of a link is specified as the start node, and which as the end node. A specified link can be traversed in either direction, dependent only on the routes specified by the timetable input. For reasons of efficiency, links should only be defined once, but the user is not confined to doing so, and it is unlikely that defining a link twice would cause any problems; however, it would clearly be wasteful of input and processing effort to do so. There is no facility at present to specify such link properties as maximum speed, curvature or gradient, the last of which is clearly ‘direction dependent’. Node-to-node travel times and average speeds are instead derived solely from the specified timetable data; this is clearly a simplification of reality, which it is planned to address in future refinements of the model.

### **Signalling System Description and Definition**

- 4.10 As noted above in para. 4.6, the decision was made to specify the signalling system separately from the track network, for reasons of simplicity and flexibility. As noted in para. 2.36, in theory, if completely conflict-free timetables were developed and adhered to, there would be little or no need for a signalling system. In reality, requirements for safety and operational efficiency usually dictate the use of some type of formal system of signalling and train control.
- 4.11 The most common signalling systems currently in use in Britain are of the three- and four-aspect coloured light type, as reflected by Arup’s model specification

(see para. 4.5, above), and the current version of the model allows the specification of either of these types. It is not possible to combine the two types in the current version of the model, although such a facility may be introduced in the future; it is also hoped to develop facilities for the simulation of the European Rail Traffic Management System/European Train Control System (ERTMS/ETCS) and other Automatic Train Protection (ATP) and Moving Block systems, in recognition of their application to the Channel Tunnel Rail Link (CTRL) and DLR, and the aspiration to introduce them more widely on the national railway network, as opportunities and funds allow (Hall, 2001, pp91-92; SPG Media Limited, 2004; Docklands Light Railway Ltd., 2002; Strategic Rail Authority, 2004). There seems to be little point in introducing facilities to simulate the behaviour of the Train Protection and Warning System (TPWS), since that would require the deliberate introduction of a facility to allow trains to pass signals at danger. The same applies with ATP in the current version of the model, and any future facility to simulate ERTMS/ETCS would automatically include ATP. Should it be required, it should also be possible to develop the means of simulating absolute block systems, with semaphore and/or two-aspect colour light signalling.

- 4.12 The data structure adopted for the purposes of defining the signalling system is summarised in Table 4.2. The first four table entries comprise general information identifying the specified type of signalling, and the base width, post height and lamp size to be used when drawing the signals on the screen. The information required for each block comprising the signalling system is then described.

**Table 4.2: Data Used to Define Signalling System**

<b>Data Item</b>	<b>Comment</b>
General System Data:	
Signalling Type	Three- or Four-Aspect in the current version of the model; options to be expanded in due course
Base Width	The width in pixels of the signal bases, i.e. the extent to which the signal posts are offset from the track on the screen – should be greater than the specified lamp radius
Post Height	The height in pixels of the signal posts, i.e. the distance between the signal base and the bottom of the lowest lamp
Lamp Radius	The radius in pixels of the signal lamps – should be less than the signal base width
For each Block in the Signalling System:	
Start Track Label	The identifying label of the track section at the start of the signal block (i.e. at the location of the signal protecting the current block, or at a point of entry to the modelled network)
Start Chainage	The chainage at the start of the block
End Track Label	The identifying label of the track section at the end of the signal block (i.e. at the location of the signal protecting the block in advance, or at a point of exit from the modelled network)
End Chainage	The chainage at the end of the block.
Overlap Length	The length of overlap provided in advance of the signal protecting the block in advance.
Signal ID	The identification label of the signal protecting the block in advance.
Entry Block	A Boolean variable (value = True or False) indicating whether or not the current block starts from an entry point to the modelled network
Exit Block	A Boolean variable indicating whether or not the current block ends at an exit point from the modelled network
Interlocking Data	The block(s) with which the current block is interlocked, if any, and the nature of the interlocking.
Protecting Block and Signal Data	The block and signal (if any) protecting the current block.

The aspiration to further expand the range of signalling types has already been discussed in this section. In the previous section, the current limitations of the model with respect to the input of positional data were noted; the same applies

to the specification of the signal dimensions, and the obvious future approach would be to relate the displayed size of the signals to the scale being used for the display of the modelled network.

- 4.13 The Start and End Track Label and Chainage data items are used to specify the section of track covered by the block in question, and thus to relate the signal and block locations to the underlying track network. The data is structured and processed in such a way that a signal is located at the end of each specified block, and the blocks thus run from signal post to signal post. This is at variance with the normal specification of signal sections, i.e. from overlap to overlap, or from overlap to signal post (Hall, 2001, p22), but the same effects are achieved by means of specifying the next data item, i.e. the overlap length. No distinction is currently made between controlled and automatic signals (Hall, 2001, p16), and all signals are treated as though they were automatic, but such a distinction is reasonably straightforward to introduce.
- 4.14 The Signal ID data item is simply a label applied to the signal at the advance end of the block, which can be used to identify and refer to both the signal and the block. The Entry Block and Exit Block data items are Boolean variables (i.e. with values either True or False) used to specify whether the block is at an entry or an exit point to or from the modelled network. Unless the network being modelled is extremely small, the values of the two data items would never both be True (although both will often/usually be False), but no check is currently made for this eventuality.
- 4.15 The interlocking issue remains to be fully addressed; the same is true for bi-directional working. The use of blocks is a simplification: to be more realistic, signals should be controlled by the status (i.e. occupied or not) of possibly several track circuits and by the position (i.e. normal or reverse) of any relevant sets of points (e.g. Nock, 1980, pp35-49). In this case, it would probably be best if the signals, track circuits and points were specified separately within the input data file(s), and their mutual dependencies then defined.

## Timetable Description and Definition

- 4.16 Having defined the infrastructure by means of the track network and signalling system input files, the train service to be applied to the network is specified by means of a timetable input file. The structure of this input file is based on the working, rather than the public, timetable format, with each train's route specified in terms of the network nodes along its path, and the times at which the train stops at, starts from or passes those nodes. At each node, the train may stop and/or start, or pass without stopping (Martin, 1999, p1288). This structure provides a detailed description of each train's routing through the network on a node-by-node basis. The data structure adopted for the purposes of specifying the timetable is summarised in Table 4.3.

**Table 4.3: Data Used to Specify the Modelled Timetable**

Data Item	Comment
General Timetable Data:	
Start Time	The time at the start of the simulated timetable
End Time	The time at the end of the simulated timetable
Number of Trains	The total number of trains in the simulation – again, not strictly necessary, but can be used to check for the consistency of the subsequent data.
For each Train in the Timetable:	
Number	The unique identifier of the train
Origin	The starting node of the train
Destination	The finishing node of the train
Start Time	The time at which the train is scheduled to leave or pass its starting node
End Time	The time at which the train is scheduled to arrive at or pass its destination node
Length	The length of the train – this influences the clearing of signals behind the train
Type	A description of the train's type
For each Node on the Train's Route:	
Name	The node number or label
Event(s)	What the train does at the node: Pass or Stop and/or Start
Time(s)	The time(s) at which the event(s) at the node is/are scheduled to occur

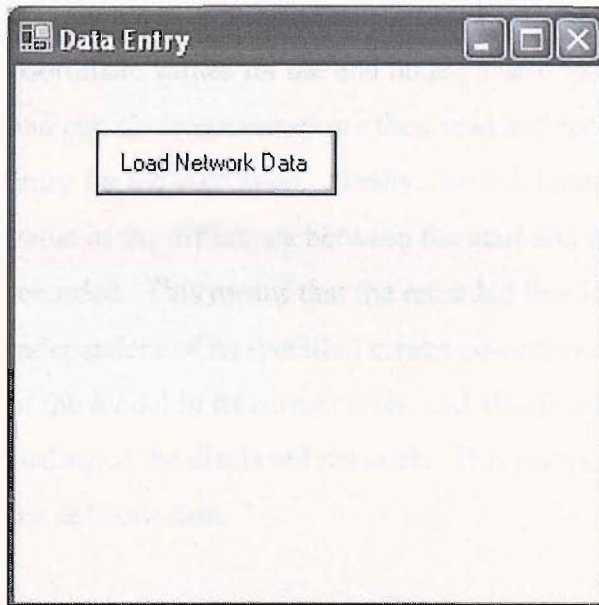
- 4.17 The timetable start and end times simply indicate the times at which the simulation starts and finishes. The simulation runs between these times, and introduces trains to the network at the scheduled intermediate times. It is possible that delays in the simulated operation of the timetable result in trains not reaching their scheduled locations by the specified simulation end time. In this case, the simulation simply stops with the modelled trains in whatever locations they have by then reached. The model could, however, be amended to give the user the option of continuing the simulation until all the trains have reached their scheduled destinations within, or have left, the modelled network.

## Reading and Preparing the Model Data

### Network Data

- 4.18 At the current stage of the model's development, when it is run, the user first sees the window shown below in Figure 4.1:

**Figure 4.1: Network Data Entry Window**



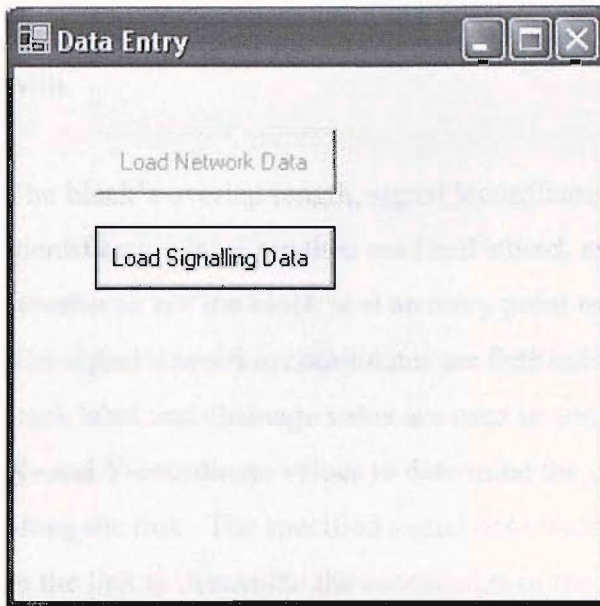
Clicking the button causes the pre-specified network data file to be loaded into the model; the name and location of the file are currently 'hard coded' into the model's network data entry routine, but an obvious improvement would be to enable the user to select the relevant file from any available drive and/or folder, using a standard File | Open dialog box.

- 4.19 The total number of nodes in the network is read from the data file, and an array of the appropriate size is set up to hold the data. The total number of links comprising the network is initially set to zero. For each node in the network, its number, X- and Y- coordinates, type and the number of nodes to which it is linked are read and stored in its array entry. The number of nodes to which it is linked (possibly none, for data input purposes, if it is at the end of a link at the edge of the network) is added to the running total of the number of links in the network. If the one or more nodes are linked to the current node, another array of the appropriate size is set up within the data record for the current node, to hold the necessary link data. For each linked node, the node number, link label and link start and end chainage values are read and stored in the appropriate array position.
- 4.20 When all the node data have been read, processed and recorded, another array is set up to hold the total recorded number of links. For each link, the start node number and X- and Y- coordinates are read and recorded, followed by the end node number. The node data array is then searched to find the X- and Y- coordinate values for the end node. These values are recorded, and the link start and end chainage values are then read and recorded from the node data array entry for the start node. Finally, the link length is calculated as the absolute value of the difference between the start and end chainage values, and is recorded. This means that the recorded link length value is completely independent of its specified screen co-ordinate values, enhancing the flexibility of the model in its current state, and also in any future version allowing the scaling of the displayed network. This completes the loading and processing of the network data.

#### Signalling Data

- 4.21 Once the 'Load Network Data' button has been clicked and the network data file has been successfully loaded, and its contents processed, the window is updated, as shown below in Figure 4.2:

**Figure 4.2: Signalling Data Entry Window**



Clicking the 'Load Signalling Data' button causes the pre-specified signalling data file to be loaded into the model; as with the network data file, the name and location of the file are currently 'hard coded' into the model's signalling data entry routine, another obvious candidate for improvement.

- 4.22 The specified signalling system type (currently restricted to either 3- or 4-aspect coloured light) is read from the data file and recorded, as are the signal base width, post height and lamp radius values. An array is set up to hold the sets of block data; each block data set is then read and processed in turn, and the resulting values are stored in the array. The block's start and end track label and chainage values are first read and stored. If the block's start and end track label values are the same, the block length is calculated as the absolute value of the difference between the chainage values. If the block's start and end track labels are different, the implication is that the block spans (at least) two different links of the network. The link data array is searched to find the links containing the two ends of the block, and the node common to the two links is found. The block length is then determined by adding the absolute values of the chainage differences between the start of the block and the 'common node' end of the link it is on, and between the end of the block and the appropriate end of its link. The current version of the model makes no allowance for situations where the (different) links containing the start and end of a block are separated by one or



more intermediate links. This might occur at a crossover, for example, and further consideration needs to be given to how such situations should be dealt with.

- 4.23 The block's overlap length, signal identification label and protecting signal identification label are then read and stored, as are the Boolean values indicating whether or not the block is at an entry point to or an exit point from the network. The signal's position coordinates are then calculated and stored: the block's end track label and chainage value are used in conjunction with the relevant link's X- and Y-coordinate values to determine the coordinates of the signal's position along the link. The specified signal base width is then projected perpendicular to the link to determine the coordinates of the foot of the signal post (all signals are located to the left of the track, as seen when facing the signal). The coordinates of the top of the signal post and of the signal lamps (three or four, as appropriate) are then calculated by projecting the appropriate distances from the base of the signal post, parallel to the link. It has been suggested by Arup that the position of the end of the overlap for each block should also be indicated on the screen; this has not yet been done, but can be achieved in a similar manner.
- 4.24 If the block is an exit point from the network, a means is required of changing its signal aspects once a train has left the network via that route. The passing of the block's overlap (or its signal, if the signal is controlled) by a train will cause the signal's aspect to change to Red. Because the block is the last one modelled on that section of the network, there is no means of determining when the train has cleared the next block, at which point the signal of the current block would revert to Yellow. This problem is addressed by creating an array of 'buffer times' for each exit block on the network, the size of the array being equivalent to the number of aspects of the signalling system in the model. When a train passes an exit block's overlap (or signal, as appropriate) the buffer times are generated and recorded (the process is described in more detail later in this Chapter), and after each buffer time increment has elapsed, the signal reverts in turn to Yellow, Double Yellow (four-aspect only) and Green, unless, of course, another train has passed in the meantime. Each exit block that is recorded in the data reading process is added to an exit block collection, for a purpose that is

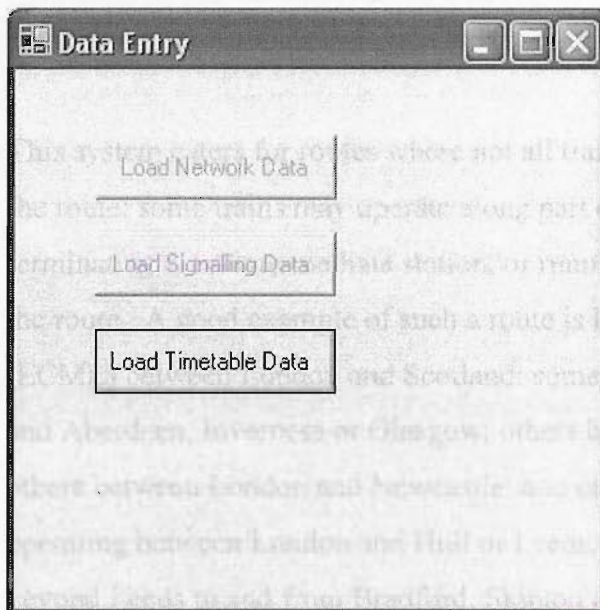
data reading process is added to an exit block collection, for a purpose that is also described later in the Chapter.

- 4.25 Finally, the aspect of the signal in each block is set to Green. This is unrealistic where interlockings occur, and is a priority for improvement.

#### Timetable Data

- 4.26 Again, once the 'Load Signalling Data' button has been clicked and the signalling data file has been successfully loaded, and its contents processed, the window is updated as shown in Figure 4.3.

**Figure 4.3: Timetable Data Entry Window**



Clicking the 'Load Timetable Data' button causes the pre-specified timetable data file to be loaded into the model; as with the previous two files, the name and location of the file are specified in the code of the current version of the model.

- 4.27 The timetable start and end times, and the number of trains specified in the timetable, are first read from the file. An array is set up, of the appropriate size to hold the specified number of trains. For each train in the timetable, its number, its scheduled start and end times and its length are read and recorded, and the numbers of nodes and links comprising its route are initialised as zero.

Arrays are set up to hold the sequence of the link data array indices of the links along the train's route, and the cumulative distance to the end of each link.

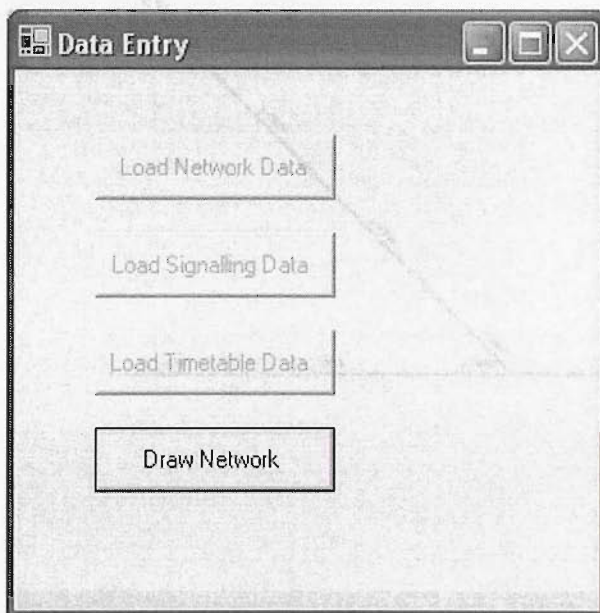
- 4.28 The remainder of the timetable data file for each train comprises a list of node labels, each with corresponding entries for arrival, departure and pass times. If the train's journey starts from the node in question (usually, but not always, the first on the list), a time will be recorded in the departure entry only; if the train's journey ends at the node in question (again, usually, but not always, the last on the list), a time will be recorded in the arrival entry only; if the train stops at an intermediate node along its route, times will be recorded in both the arrival and departure entries; if the train passes a node without stopping, the appropriate time will be recorded in the pass entry; and, finally, if a listed node is not on the route of the train, no time is recorded.
- 4.29 This system caters for routes where not all trains serve or pass all stations along the route: some trains may operate along part of the route only, either terminating at an intermediate station, or running to a station off the main line of the route. A good example of such a route is Britain's East Coast Main Line (ECML) between London and Scotland: some trains operate between London and Aberdeen, Inverness or Glasgow; others between London and Edinburgh; others between London and Newcastle; and others divert from the main route, operating between London and Hull or Leeds, with some services extending beyond Leeds to and from Bradford, Skipton or Harrogate. A list of all the nodes along a particular route can thus be used as a timetable template, with times being specified only for those nodes which a particular train serves or passes.
- 4.30 As the program reads through the list of nodes for each train in the timetable data file, the node label, event type (arrival, departure or pass) and time (hh:mm:ss) are recorded for each node with at least one specified time. The time is then converted into seconds, i.e. the number of seconds between midnight and the specified time (the calculation makes no allowance for trains whose journey starts before midnight and finishes afterwards, another candidate for improvement). The number of links comprising the route is increased by one

if the train arrives at or passes the node in question, but is not incremented if the train departs from the node, since the train is then either starting its journey with no links traversed as yet, or is resuming its journey from the end of a link that has already been counted.

- 4.31 Unlike the number of links comprising the route, the number of nodes along the route is incremented for every recorded ‘event’; at intermediate stops, where both an arrival and a departure time are recorded, the number of nodes is incremented by two, with a different time and event being recorded for each, although the two node labels are the same. For each event along the route, the cumulative distance is calculated by adding the absolute difference between the chainage values at the current and previous nodes to the previous cumulative distance value. For intermediate stops, the calculated chainage difference between the arrival and departure events is zero, so the same cumulative distance value is recorded for both events. This process results in the recording of the routes and timings for each train, ready for use in the simulation process. These calculations could be done during the simulation, but doing them at the data loading stage reduces the amount of computation to be performed at the simulation stage, enabling the simulation to run more quickly and efficiently.
- 4.32 Having established each train’s route and scheduled timings through the track network, a similar exercise is performed to establish the corresponding routings through the signalling system. The number and sequence of signal blocks on each train’s route is determined, together with the cumulative distance to each signal and overlap. The information recorded is used during the simulation to determine which block(s) each train is occupying, by comparing its elapsed distance with the various cumulative values, and thus to set the signal aspects appropriately. Again, performing these preliminary calculations at the data loading stage has the effect of reducing the extent of the computation required during the simulation process. An additional potential use of the pre-determined routes through the signalling system is for the ‘setting up’ of routes ahead of trains, as is done on parts of the British railway system by Automatic Route Setting (ARS) technology, as mentioned in Chapter 2.

- 4.33 It should be noted that the routeing of the timetabled trains is straightforward, in that the trains simply follow the scheduled routes. Things become somewhat more complicated under disrupted operating conditions, when trains may have to deviate from their specified routeings. Some trains may not be allowed to use certain routes, due to weight and/or clearance restrictions, and other constraints include drivers' route knowledge, and the availability or otherwise of bi-directional signalling. At a more detailed modelling level, a means is required of preventing the simulation from routeing trains in directions that could not be achieved in reality: for example, the structure of the link and node data might indicate to the model that it is feasible to route a train towards a trailing set of points, and then through an acute angle onto the converging track, travelling in the opposite direction, while this would clearly be impossible in reality, without reversing the train.
- 4.34 Once the timetable data file has been successfully loaded, and its contents processed, the Data Entry window is updated once more, as shown in Figure 4.4.

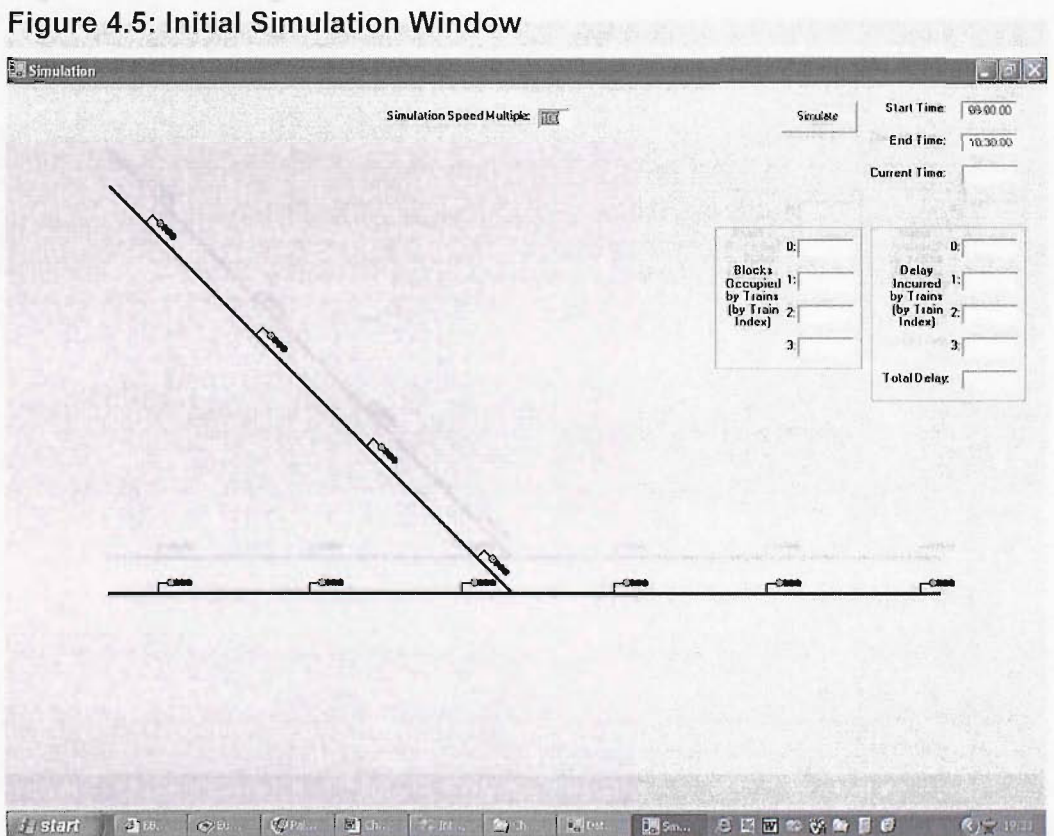
**Figure 4.4: Simulation Preparation Window**



Clicking on the 'Draw Network' button results in the simulation window being displayed, as shown in Figure 4.5.

4.35 The Simulation window displays the track network being modelled, together with the specified signals (all initially set to Green aspects, as already noted). The 'Simulation Speed Multiple' text box allows the user to specify the speed at which the simulation runs: in the example shown, the multiple of 100 will result in the specified timetable duration of 01:30:00 being simulated in 54 seconds ( $= 90/100 = 0.9$  minutes), or, at least, as quickly as the computer can manage. The timetable duration can be determined from the difference between the timetable start and end times, displayed in the text boxes in the top right corner of the window; when the simulation is running, the current simulated time is displayed in the third text box, captioned 'Current Time'. The other text boxes shown are used for the purpose of testing the program, and are unlikely, with the possible exception of the 'Total Delay' text box, to be retained in working versions of the model.

Figure 4.5: Running Simulation Window



## Running the Model Simulation Window

- 4.36 The model is run by clicking the 'Simulate' button, which causes the window's appearance to change slightly, as shown in Figure 4.6, below. No trains have yet been introduced to the network, but the 'Current Time' text box displays the simulated time, and a 'Pause' button has been added to the window, enabling the user to temporarily halt the simulation, for example to check some aspect of the model run. The effect on the window of clicking the 'Pause' button is shown in Figure 4.7, below: the simulation can now be resumed by clicking the 'Resume' button, or reset to its initial state by clicking the 'Reset' button. The only current means of terminating the simulation is by clicking the red 'Close' button in the top right corner of the window, but a dedicated 'Exit' button and/or menu option can easily be provided.

Figure 4.6: Running Simulation Window

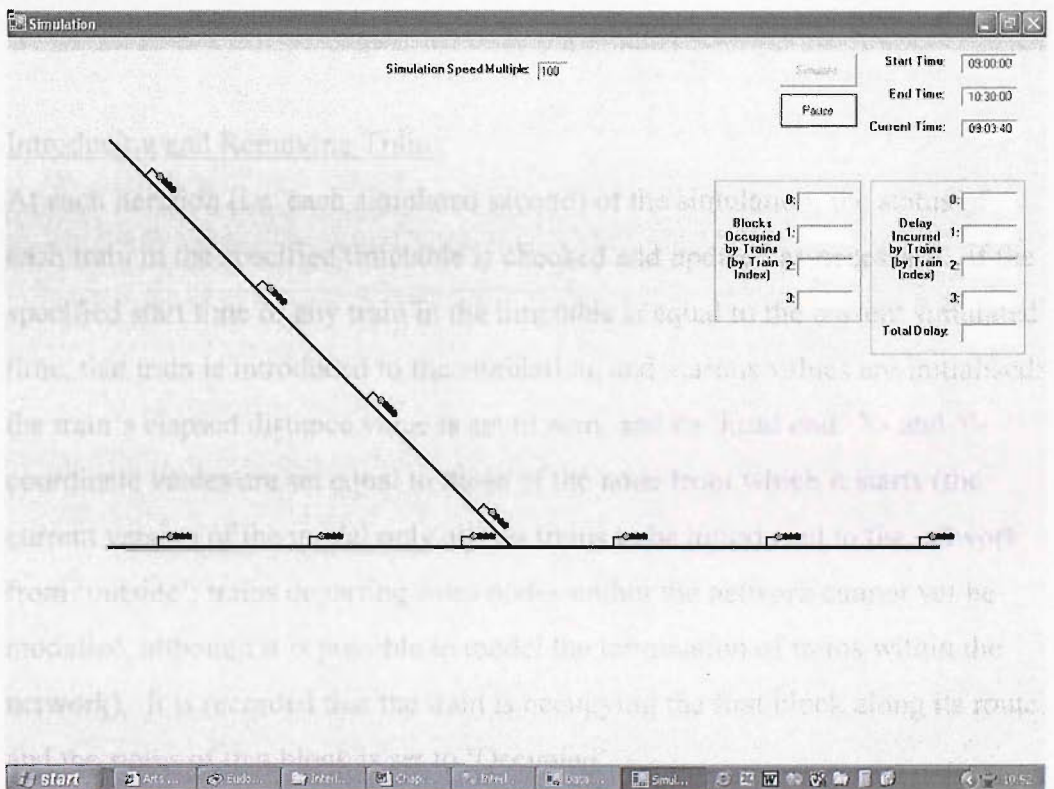
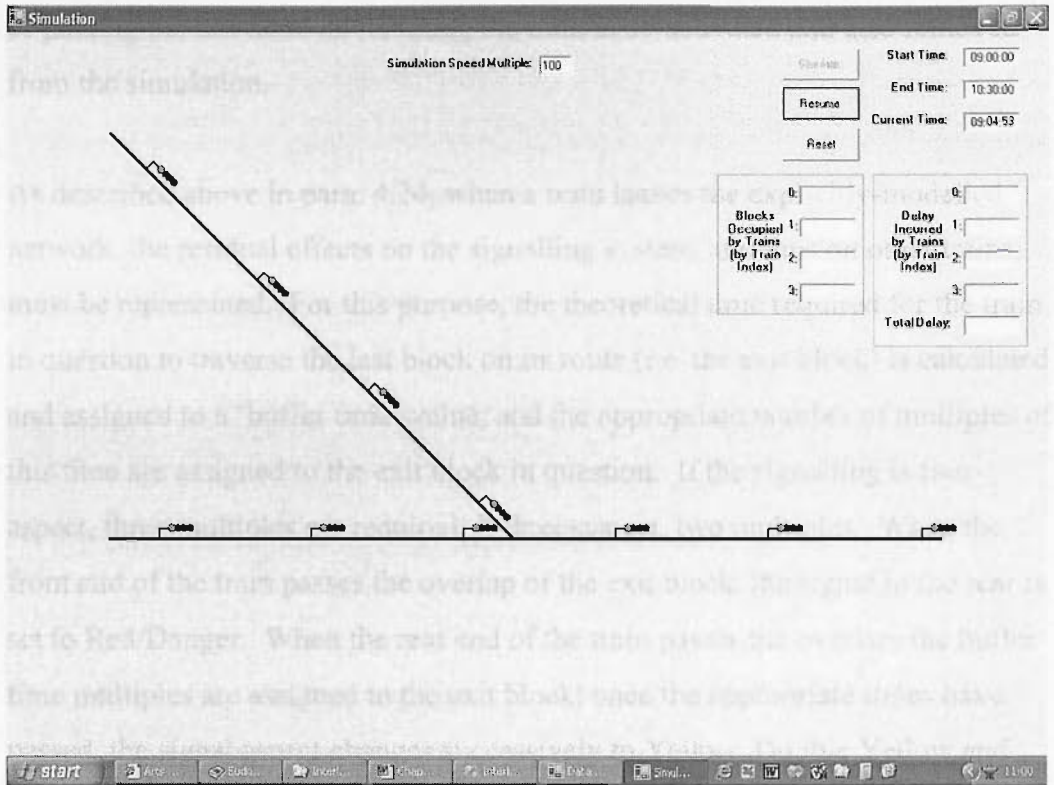


Figure 4.7: Paused Simulation Window



#### Introducing and Removing Trains

- 4.37 At each iteration (i.e. each simulated second) of the simulation, the status of each train in the specified timetable is checked and updated as necessary. If the specified start time of any train in the timetable is equal to the current simulated time, that train is introduced to the simulation, and various values are initialised: the train's elapsed distance value is set to zero, and its 'head end' X- and Y-coordinate values are set equal to those of the node from which it starts (the current version of the model only allows trains to be introduced to the network from 'outside'; trains departing from nodes within the network cannot yet be modelled, although it is possible to model the termination of trains within the network). It is recorded that the train is occupying the first block along its route, and the status of that block is set to 'Occupied'.
- 4.38 If a train arrives at its terminal destination within the modelled network, it is deactivated, but is retained in the simulation and remains visible on the display. The status of the signalling system also reflects its presence (i.e. the status of the block holding it remains 'Occupied', and the signal in its rear remains at



Red/Danger). If, on the other hand, a train leaves the network by departing from or passing the last node on its route, the train is de-activated and also removed from the simulation.

- 4.39 As described above in para. 4.24, when a train leaves the explicitly-modelled network, the residual effects on the signalling system, and thus on other trains, must be represented. For this purpose, the theoretical time required for the train in question to traverse the last block on its route (i.e. the exit block) is calculated and assigned to a 'buffer time' value, and the appropriate number of multiples of this time are assigned to the exit block in question. If the signalling is four-aspect, three multiples are required; if three-aspect, two multiples. When the front end of the train passes the overlap of the exit block, the signal in the rear is set to Red/Danger. When the rear end of the train passes the overlap, the buffer time multiples are assigned to the exit block; once the appropriate times have passed, the signal aspect changes successively to Yellow, Double Yellow and Green (four-aspect) or Yellow and Green (three-aspect). If another train leaves the network from the same exit block at any time between the first and last buffer times, the signal aspects triggered by the first set of buffer times are overridden by the passage of the subsequent train, and a new set of buffer times is generated. This method makes the assumption that the next three (four-aspect) or two (three-aspect) blocks beyond the explicitly-modelled network boundary can be traversed in the same length of time as the exit block. An alternative approach would be to model the necessary number of blocks beyond the boundary without representing them on the screen or in any other output.

#### Updating Train Positions

- 4.40 For all active trains (i.e. those that have started their journeys but not yet reached their destinations), their elapsed distances are updated at each iteration of the simulation. In the current version of the model, a train's speed between nodes is assumed to be a constant value, calculated by dividing the distance between the nodes by the specified travel time between them. The distance increment for each second of the simulation will therefore be equal to the calculated average inter-nodal speed (m/s), unless the train has to stop at a signal set to Red/Danger. This calculation of the distance increment is clearly a simplification, but one that

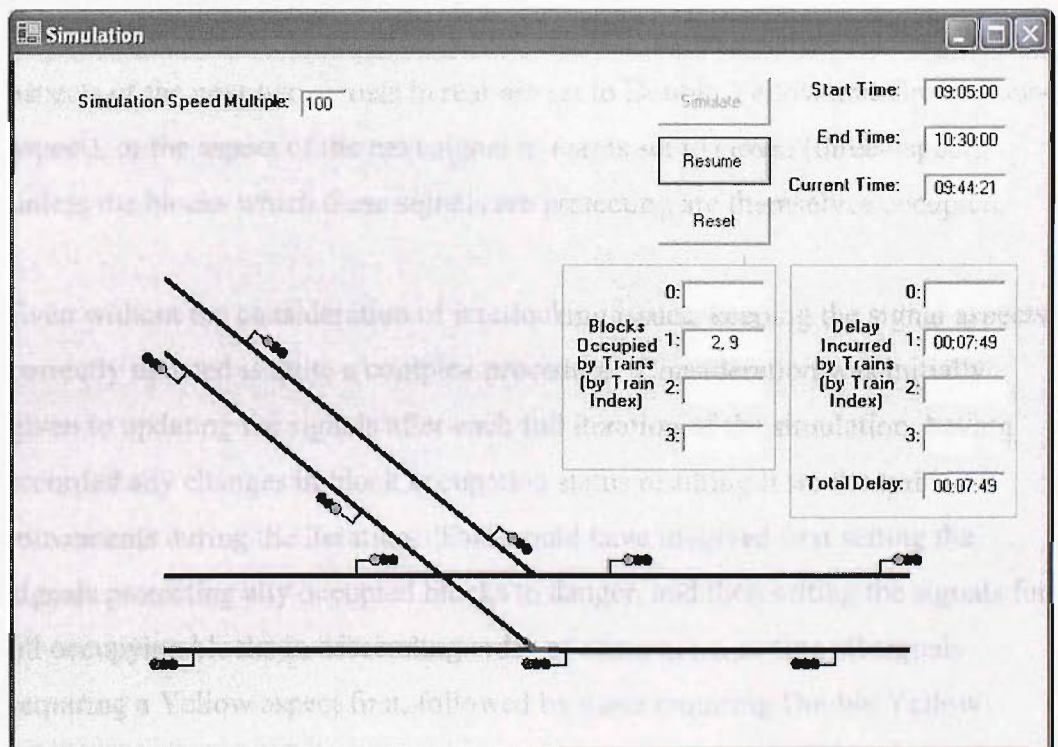
is apparently used by some of the commercial models on the market, and the distance increment calculation can be replaced (possibly as an option, to be selected according to the level of detail required in the simulation) in future updates of the model by calculations based on the linear or differential equations of motion. Use of the linear, Newtonian equations would be suitable for the modelling of Light Rail operations, while use of the differential equations would be more appropriate for the realistic simulation of Heavy Rail operations, although it would also be much more computationally demanding. In the early stages of the research work, it was intended to incorporate the algorithms developed for the spreadsheet model in the more general model. However, in the course of the model development it became apparent that the differences in the structures and objectives of the two models would make this impracticable.

- 4.41 As noted in the preceding paragraph, if a train encounters a signal set to Danger, it stops. This occurs if the elapsed distance plus the calculated distance increment is greater than the distance to the next signal, and the signal's aspect is Red. In such a case, the calculated distance increment is replaced by the distance to the next signal, so that the train does not pass it. Because this approach is adopted, signals whose aspects are set to Yellow or Double Yellow have no effect on the behaviour of the simulated trains, unlike their real-life counterparts, where train drivers need to make a brake application when a signal displaying either of these aspects is encountered, in order to bring the train safely to a stop without incurring a SPAD (Signal Passed At Danger). The display of the Yellow and Double Yellow aspects enhances the appearance of the model display, however, and the aspects may be used directly to trigger simulated train deceleration in later versions of the model, when train acceleration and deceleration are explicitly modelled.
- 4.42 If a train's elapsed distance plus the distance increment equals or exceeds the distance to the next node on the route, the 'next node' value for the train is updated to the label corresponding to node on the route after the one just reached or passed. If the train is scheduled to arrive (i.e. stop) at the now-current node, and the calculated distance increment results in the train overshooting the node, the distance is amended appropriately. Additionally, if the node is not the last

on the route (i.e. the train is making an intermediate stop), the actual time of arrival is compared with the scheduled arrival time. If the actual time is later than the scheduled value, the scheduled departure time is incremented by the difference, so that the train's scheduled dwell time is maintained. A preferable, more flexible approach would be to specify a minimum dwell time for the stop, and to replace the scheduled departure time with the actual arrival time plus the minimum dwell time, if that were later than the scheduled departure time. Once the train's scheduled/amended departure time is reached, its journey continues.

- 4.43 As noted in the previous paragraph, a train's 'next node' value is updated as it passes nodes along its route. This procedure is carried out for both the front and the rear of the train, and the train can then be represented on the screen by
- 4.44 comparing the elapsed distance and the (elapsed distance – train length) values with the cumulative distances to the nodes between which the front and the back of the train are located. Interpolating between the appropriate node coordinates produces the coordinates of the ends of the train, between which a line is drawn, representing the train. This is straightforward in situations where the entire train

**Figure 4.8: Drawing of Trains on Screen**



is located between the same two nodes, or where the train's length occupies two or more links on the same straight line. The situation becomes more complicated, however, when these conditions do not apply, as shown in Figure 4.8 (note: the signal aspects shown in the Figure are incorrect, in that no interlocking conditions have been applied). In situations where a train is 'going round a corner' at a junction or a change in track alignment, a more sophisticated approach is required. To meet this need, the 'next node' values for the front and rear of the train are compared: if they are not the same, the train is 'straddling' one or more nodes, and the line representing the train is drawn between the front and back coordinates of the train, via the straddled node(s).

#### Updating the Signalling Display

- 4.44 As noted in para. 4.41 above, the simulated trains respond to signals by stopping if a Red/Danger aspect is displayed. Conversely, and obviously, the signals in turn respond to the passage of trains through the modelled network and signalling system. When the front of a train passes the overlap of a signal block, the block in advance (i.e. the one which the train now enters) becomes occupied, and the aspect of the signal in the rear is set to Red/Danger to protect the train. Similarly, as the rear of a train passes an overlap, the block in its rear ceases to be occupied, and the aspect of the block's protecting signal is set to Yellow. The aspects of the next two signals in rear are set to Double Yellow and Green (four-aspect), or the aspect of the next signal in rear is set to Green (three-aspect), unless the blocks which these signals are protecting are themselves occupied.
- 4.45 Even without the consideration of interlocking issues, keeping the signal aspects correctly updated is quite a complex procedure. Consideration was initially given to updating the signals after each full iteration of the simulation, having recorded any changes in block occupation status resulting from the train movements during the iteration. This would have involved first setting the signals protecting any occupied blocks to danger, and then setting the signals for all occupying blocks in descending order of caution, i.e. setting all signals requiring a Yellow aspect first, followed by those requiring Double Yellow (four-aspect only) and then those requiring a Green aspect. This would have involved fairly extensive computation, which would have been quite inefficient

if, say, a small number of trains were occupying a large network, thus requiring the updating of only a small number of signals from iteration to iteration of the simulation. More seriously, updating the signals only after all train movements for the iteration were complete could result in a train passing a signal whose aspect should already have been set to Danger as the result of the movement of another train (this is a potential consequence of the train movements being considered sequentially within each iteration).

- 4.46 For all these reasons, it was decided to update the signals on a train-by-train basis, as happens in reality, with each train ‘taking care of’ the signals whose aspects it is affecting at any given time. To this end, each train in the simulation has a facility for holding the identities of the blocks it is currently occupying, and, separately, the identities of those blocks whose signals are providing it with protection from following trains. As the front of a train occupies successive blocks along its route, those blocks are added to its ‘occupied blocks collection’, the aspects of their protecting signals are set to Red/Danger, and the blocks in rear of the protecting signals (to which the protecting signals are assigned in the input data) are added to its ‘protecting blocks collection’. As the tail of the train clears blocks along the route, they are removed from the occupied blocks collection, the aspects of their protecting signals are set to Yellow, and those of the preceding protecting blocks are set to Double Yellow and/or Green. Once a protecting block’s signal aspect reverts to Green, it is removed from the train’s protecting block collection, since it no longer protects the train by instructing following trains to ‘Stop’ or ‘Proceed With Caution’.
- 4.47 Blocks can also be removed from a train’s protecting block collection by the movement of another train. If, in the interval since the previous iteration, a block in a train’s protecting block collection whose own protecting signal aspect was Yellow or Double Yellow has become occupied by a following train, then its protecting signal aspect will change to Red, to protect the occupying train. The block containing the Red signal will now no longer be protecting the first train, and will therefore be removed from its protecting block collection.

4.48 At the end of each iteration of the simulation, the aspects of any signals which have been set to protect trains that have left the network are updated as necessary (see paragraphs 4.24 and 4.39, above). Since the updating of these signals is time-based, there is no point in updating them more than once per iteration. As described in the preceding paragraph, the occupation of the protecting blocks by following trains may result in the removal of their (i.e. the following trains') protecting blocks from the protecting block collections of trains that have left the network.

#### Calculation of Delay

4.49 As noted in Chapters 2 and 3, delay is a fundamental measure of railway performance, and can be used to assess the quality of a proposed timetable or the likely effects of proposed changes to infrastructure and/or train services. It is also a valuable means of comparing the outcomes of different responses to disruptive incidents. It is therefore essential to the aims of this research project that the model can calculate any delay incurred by simulated trains.

4.50 Because train performance is not explicitly modelled in the simulation, and trains run exactly according to the specified timetable, the only way in which they can incur delay is by having to stop at a signal set to Danger. Similarly, because there is no 'slack' in the timetable, trains cannot run ahead of time, thus incurring negative delay, or reducing the amount of any delay that has already been incurred.

4.51 The delay incurred by a train at any location on its route can be determined from the difference between the current (simulation) time and the time at which the train is scheduled to be at that location. If the train is scheduled to arrive at or to pass the next node on its route, the remaining journey time from its current location to that node is calculated and then subtracted from the scheduled time of arrival at or passing of the node. This gives the scheduled time at the train's current location, which is subtracted from the current time to yield the current delay value. If the train's next scheduled event is departure from a node (i.e. it is already at the node in question), then if the current time is later than the scheduled/amended departure time (see para. 4.42), the delay is calculated as the

difference between the two. If the current time is earlier than or equal to the departure time, the delay is zero. The individual train delays are calculated sequentially (i.e. train-by-train) within each iteration, and the individual values are summed at the end of each iteration to produce a total delay value.

#### Ending the Simulation

- 4.52 As noted in para. 4.17, the simulation ends once the specified timetable end time is reached, although this could be amended to enable any late-running trains to reach their destinations. The simulation window is then closed by clicking on the standard Windows ‘Close’ button, as described in para. 4.36, although, as noted in the same location, this could also be amended and improved.

#### **Further Work**

- 4.53 The work to date has concentrated on the fundamentals of network, signalling and timetable data structures, the routing of trains through the track network and their interactions with the signalling system, the visual representation of the network, signals and trains, and the calculation of any incurred delay times.
- 4.54 In its current state, the model is much less ‘complete’ than the spreadsheet model described in Chapter 3, and has extensive possibilities for improvement. These include:
- Visual representation of bridges, tunnels, stations, grade separation of junctions, etc.
  - The automation of coordinate conversion, and the associated introduction of a facility to scale a network according to the extent of the model, including signal dimensions, etc.
  - The provision of a facility to select input files from dialog boxes.
  - The provision of facilities to specify link maximum speeds, gradients, radii of curvature, and rolling stock/train type restrictions.

- The introduction of additional node types: ‘graphical’ nodes to represent the convergence and divergence of parallel tracks; and nodes to specify change of track label and/or chainage, gradient and curvature.
- The provision of a facility to specify movements which cannot take place at junctions, for use with re-routeing options.
- The simulation of interlocking of signals and points.
- The introduction of facilities to simulate other signalling systems, including ERTMS/ETCS; to specify signals and track circuits, rather than signalling blocks; to allow the simulation of bidirectional signalling; to distinguish between automatic and controlled signals; and to automatically specify default overlap lengths (Hall, 2001, p22; Nock, 1980, p5).
- Other miscellaneous improvements include the extension of simulation times as necessary to allow all trains to reach their scheduled destinations or to leave the modelled network; the automatic calculation of block lengths where the (different) links containing the start and end of a block are separated by one or more other link(s); visual indication of overlap ends; allowance for timetables which include midnight; development of the ability to start trains from within the modelled network; and the specification of minimum dwell times at scheduled stops.



## **5.0 DISRUPTION MANAGEMENT**

- 5.1 Having introduced the research topic in Chapters 1 and 2, and described the requirements, underlying principles and development of two models in Chapters 3 and 4, this Chapter now considers the disruption of railway operations at a greater level of detail. It examines the consequences of disruptive incidents and their effects on route capacity, the available options for dealing with them, the objectives underlining disruption management, practical and theoretical approaches to achieving those objectives, and, finally, it considers possible alternative approaches and proposes a revised framework to be adopted when dealing with disruption.
- 5.2 Following this introduction, the related issues of capacity and disruption are first examined. The various practical ‘tools’ available for addressing disruptive incidents are then considered. Next, the principles and methods currently in use in the British railway industry are summarised. A general framework for dealing with disruptive incidents is then proposed; one that has similarities to the approaches currently in use, but also some significant differences. Finally, some potentially useful areas for further work are identified.

### **Capacity, Disruption and Delay**

- 5.3 Railway capacity is a complex quantity. Pachl (2002, p215) defines the “capacity of a line [as] the number of trains which may be operated through [it]”, but observes (p137) that “it is not possible to determine one single measure for the capacity of a whole railway network or larger parts of a network” and that “the possible exploitation of one part of the network depends not only on the theoretical capacity of this part but also on the capacity of the adjacent parts.” Harris (2003a, p1) notes that “railway systems have a fixed capacity, even if this is difficult to define because it may be different along a length of line rather than at a point.” Elsewhere (2003b, p1), he states that

*the theoretical capacity of a section of railway line is given by Scott's formula:*

$$C = (24 * 60) / (T + t)$$

*where*

*C = sectional capacity in number of trains per day in either direction;*

*T = time taken in minutes by slowest train to cover longest block section;*

*t = time taken in mins. to close the line, obtain line clear and start next train*

and that

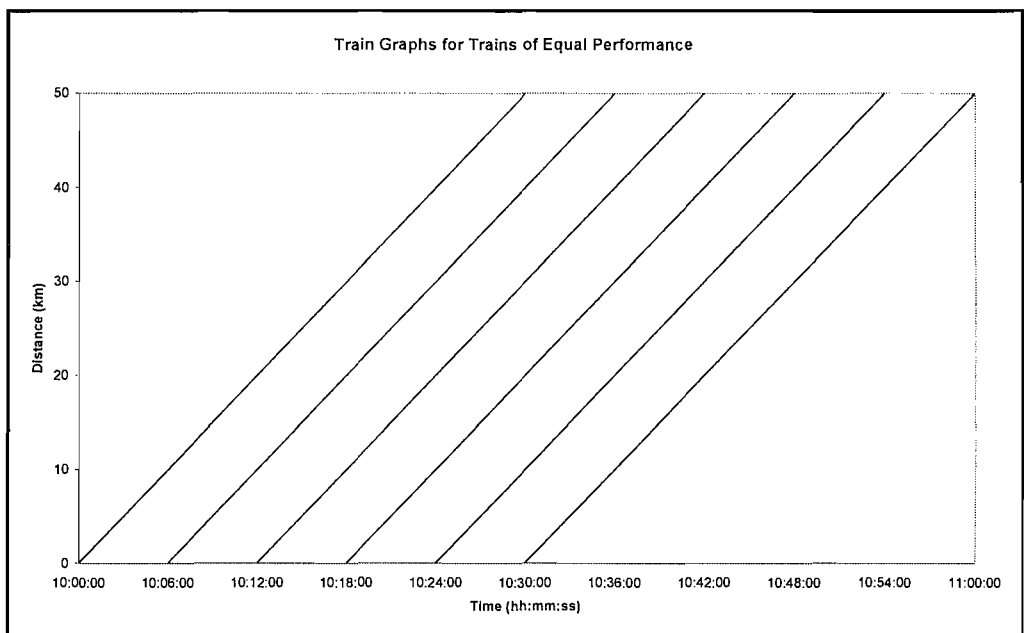
*in practice, capacity used is typically 70% of theoretical capacity, permitting some perturbation due to disruption to be absorbed, and leaving space for a small number of trains to be added to the timetable at the last minute.*

- 5.4 Given the relationship between *C* and *T* in the formula in the preceding paragraph, it can be seen that theoretical railway capacity depends on line speed, train performance and signal spacing. The capacity of a route will obviously depend on the number of lines comprising the route, but, as implied above, capacity also depends on the mix of services using the route: the effect that this can have is illustrated in the SRA's Network Utilisation Strategy (NUS) document (Strategic Rail Authority, 2003, p16):

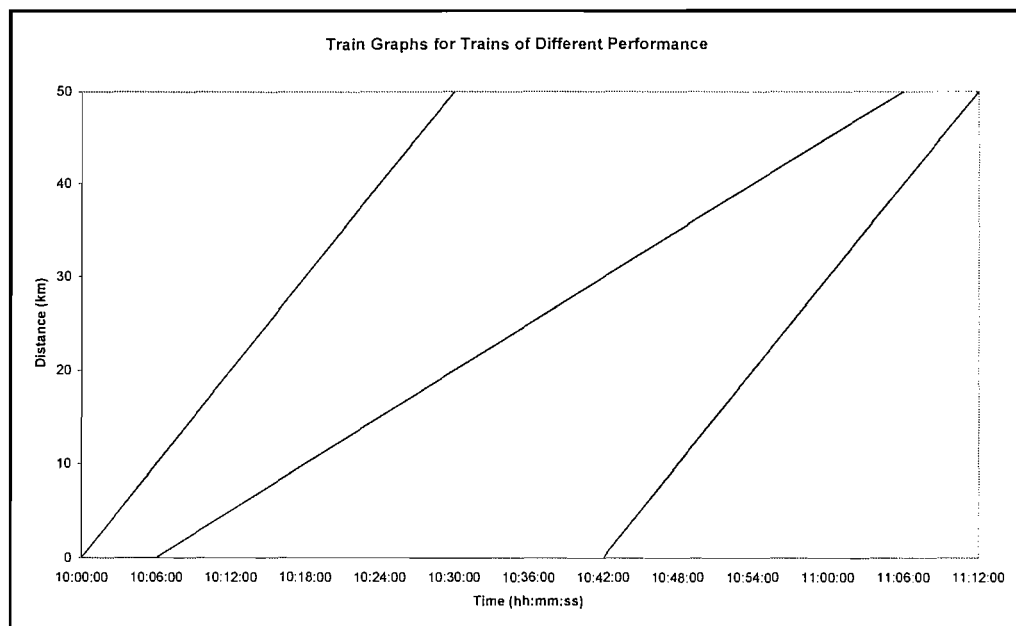
*The pattern and mix of train services is one of the major factors determining the real capacity of any route. For example, the theoretical capacity of the line between Leeds and York is 20 trains per hour. In practice, after providing a performance cushion, only 15 same-speed trains can be run. But, due to the pattern of mixed speed services currently in the timetable, and the impacts of trains that join, leave or cross the route, typically only five trains are actually run, even in the peak.*

The effects on line capacity of shared use by trains of different speeds are shown graphically in Figures 5.1 and 5.2, using an example and data from Ford and Haydock (1992, pp121, 122). The Figures represent theoretical train graphs for a 50km line without passing loops. In the first, express trains travelling at an average speed of 100km/h are operating at 6-minute intervals (headways), allowing the operation of six trains in a 30-minute period. In the second, an express train at 10:00:00 is followed by a slower train, with an average speed of 50 km/h, at 10:06:00. This prevents the next express train from operating until 10:42:00, restricting the capacity of the line to three trains in a 42-minute period. Similarly, but perhaps counter-intuitively, the operation of a fast train between two slower services has a similar effect, again causing a reduction in line capacity.

**Figure 5.1: Train Graphs for Trains of Equal Performance**



**Figure 5.2: Train Graphs for Trains of Different Performance**



5.5 As noted above, to allow the absorption of perturbations and to provide some flexibility for the operation of additional services, full theoretical route capacity is not normally used. It can be seen that the initial, practical capacity value of 15 trains per hour quoted by the SRA for the line between Leeds and York, equal to 75% of the theoretical capacity, corresponds quite closely to Harris’ value of 70%. Elsewhere (p9), the NUS document refers to the Capacity Utilisation Index (CUI) developed by Network Rail, describing it as “a measure of the level of congestion on the network [which is] essentially the proportion of theoretical train paths that are actually used.” The CUI for a route

*reflects both the number and mix of services. For example, a metro style timetable (where trains have identical speeds and stopping patterns) might allow 20 trains per hour at a CUI level of 67%, whereas a timetable over the same route accommodating trains with very diverse speeds and stopping patterns might reach the same CUI level with just eight trains per hour.*

5.6 The proportion of the theoretical capacity that is used has a strong, direct influence on the delay that is likely to occur on a line. Pachl (2002, p139) observes that

*the capacity of a line can be described in [the] form of a waiting time diagram [, in which] the average waiting time per train is shown as a function of the traffic flow (trains per unit of time).*

This diagram is similar in form to Figure 5.3 below, showing a sharply-rising and non-linear relationship between the average waiting time per train and the traffic flow as a proportion of line capacity. Pachl goes on to identify two categories of waiting time:

- *the scheduled waiting times ... due to scheduled passing and overtaking operations.*
- *the delays in current operations [i.e. unscheduled operational delays].*

He describes the waiting time as

*a measure of the quality of the operation. The waiting time curve approaches a vertical tangent which is the maximum capacity of the line. This is the maximum output of the line regardless of quality. When the input of the line exceeds the maximum capacity ... an increasing queue of waiting trains [will occur].*

5.7 The SRA's NUS (p9) distinguishes between two categories of delay:

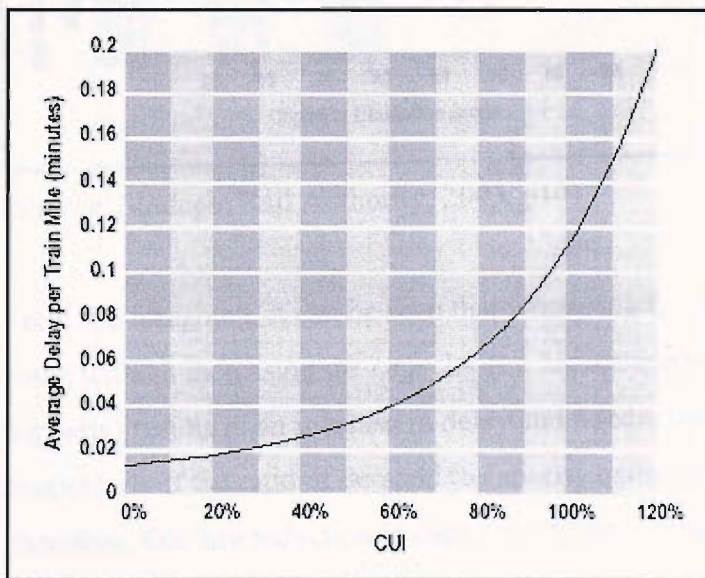
- *'Primary' delays – resulting directly from a problem with train operations, i.e. infrastructure failures, train failures, staff shortages, etc.*

- *'Reactionary' delays – where trains are held up because of knock-on effects from a previous primary delay, but not due directly to the initial cause. For example, a train delayed due to a signal failure on departure may subsequently cause delays to other trains throughout its journey. The more congested the network, the greater the likely level of reactionary delays that results from a primary incident.*

Harris (1992, p132) puts it succinctly: “the busier a railway is, the more likely it is that one failure will cause problems for other services.” The NUS (p9) quotes Network Rail data showing that 41% of delays in 2002/3 were Primary, and 61% were Reactionary.

5.8 The SRA examined the relationship between CUI and reactionary delay minutes for a range of services on different routes at different times of day. The results are shown in Figure 5.3.

**Figure 5.3: Reactionary Delay vs. CUI**



(Source: Strategic Rail Authority, 2003, p10)

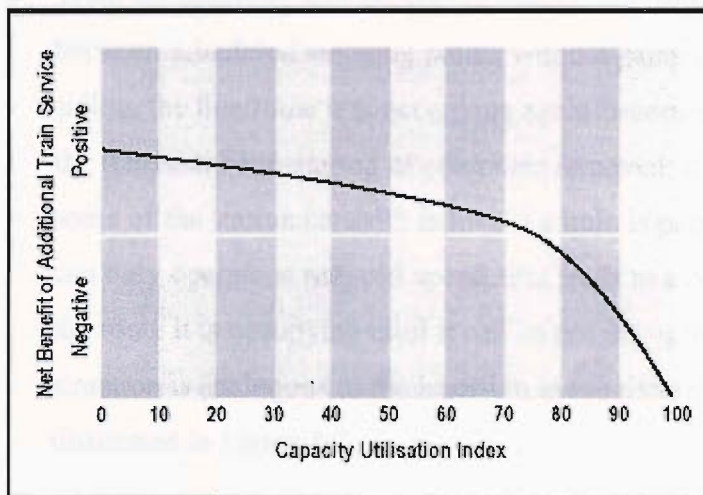
These results prompted further analysis by the SRA of the relationship between the benefits arising from the introduction of new train services on a route, and the disbenefits resulting from deterioration in route performance. The results of

this analysis are shown in Figure 5.4 below, and (Strategic Rail Authority, 2003, p10) indicate that

*for an 'average' train, evaluated in accordance with the SRA's appraisal criteria, ... costs, including performance impacts, outweigh benefits ... when the CUI is [greater than or equal to] approximately 75%.*

This proportion of theoretical capacity again agrees closely with the values quoted in paragraphs 5.3 and 5.5, above.

**Figure 5.4: Net Benefits of Additional Train Services**



(Source: Strategic Rail Authority, 2003, p10)

- 5.9 From the foregoing, it can be seen that increasing the number of services on a route without increasing the capacity (i.e. increasing the ratio of demand to capacity) results in an increase in delay and a reduction in service quality, particularly if the ratio of demand to capacity exceeds about 70-75%. It follows, therefore, that any reduction in capacity will have a similar effect, since the ratio of demand to capacity will again be increased, unless services are removed from the route. Any disruption which causes a reduction in capacity will therefore reduce the number of trains per unit time that can traverse the infrastructure in question.

5.10 The three main causes of such disruptive events are listed in para. 5.7 above, as the three main causes of primary delay identified by the SRA:

- Infrastructure failures include broken rails, point and signal failures, and route blockages by ‘Acts of God’ or, for example, the blocking of lines by road vehicles which have left the road, or at level crossings. Track failures or blockages result in the complete temporary loss to the route and/or network of the capacity provided by the affected link(s). A signal failure, on the other hand, may allow continued working under carefully controlled conditions (Hall, 2001, pp32, 54, 91), with a partial, rather than total, loss of capacity on the affected link/route/section of the network.
- Train failures may also be total or partial: if a train breaks down completely between scheduled stopping points without gaining access to a loop or siding, the line/route it is occupying again becomes completely blocked until the train can be re-started or otherwise removed; if a train fails in a station, some of the station capacity is lost; if a train is partially disabled, so that it can only operate at reduced speed, this leads to a partial loss of capacity on the route it is occupying until it can be put into a loop or siding. The latter situation is analogous to the insertion in the timetable of a slow train, as illustrated in Figure 5.2.
- Staff shortages can cause or exacerbate disruptions in various ways. If a relief train crew is unavailable, a train may be delayed at a station or crew changeover point, temporarily reducing the capacity at that location. A shortage of signalling staff may reduce the rate at which trains can be routed through the section of the network controlled by the signalling centre in question. A shortage of maintenance/repair staff may result in the prolonging of a train or infrastructure failure, thus extending the period during which capacity is reduced.



## Responses to Disruptive Incidents

5.11 As noted in Chapter 2, investment in reliability and capacity can be employed to reduce the likelihood and consequences, respectively, of disruptive events. This Chapter, however, is concerned more with the immediate, short-term responses to disruptive events when they occur. The SRA's NUS (p12) identifies six approaches to improving the use of existing capacity in situations where the CUI value exceeds the recommended maximum:

- *Increase load factors (where trains are not overcrowded already).*
- *Increase train lengths.*
- *Improve the use of available train paths.*
- *Adjust pattern and mix of services to make better use of capacity.*
- *Improve robustness of timetables.*
- *Improve the regulation of trains.*

Under disrupted operating conditions, it may well be necessary temporarily to increase load factors, even if trains are already crowded. The next four measures all come into the category of long-term responses, but the final measure has considerable potential as an immediate response. The NUS document (pp12-17) elaborates on the first five measures, but has nothing further to say about the sixth, other than to suggest the revisiting of “prioritisation rules, class regulation practices and use of passing facilities by passenger services”, as already noted in para. 2.50 of Chapter 2.

5.12 When disruptive incidents occur, there is typically an aspiration to maintain the best possible level of service in the circumstances, although the emphasis may vary. When passenger services are delayed, overcrowding may occur at busy commuter or, especially, Metro/Underground stations, endangering passengers. Overcrowding may also occur on trains, as delayed services pick up passengers who would normally have travelled on a later-scheduled train. In such situations, it may be desirable and necessary to alter train stopping patterns and

timings to alleviate crowding on trains and/or at stations (Goodman and Takagi, 2004, p767). Discussions with staff at two London Underground control centres during visits in November 2002 indicated that passenger safety is the overwhelming priority in the case of service disruptions. If a line becomes blocked, or in the case of other serious delays, the overriding objective is to get trains into stations so that passengers can be evacuated from the system; maintaining services is very much a secondary issue.

- 5.13 White (2003a, p2), referring to heavy rail operations in North America, places more emphasis on maintaining services:

***The Trains Must Go On***

*Train dispatchers, unlike air traffic controllers, must handle all traffic as it occurs, regardless of congestion or weather. Rerouting of traffic occurs only in instances of impassable blockage or destruction of the track by accident or weather. When snow slows traffic to a crawl, when part of the available trackage is rendered unusable by accident or weather or immediately upon partial restoration after complete blockage or destruction, traffic will be operated normally and it is up to the train dispatcher to handle normal traffic with the reduced facilities while concurrently allowing the maintenance or rerailing work necessary to restore the trackage completely. The resulting congestion itself causes further reduction of track capacity for which the dispatcher must compensate while operating as closely to regular schedules as possible and ensuring that an absolute minimum number of trains must be relieved because of hours of service law. Likewise, such incidents as well as accidents that do not involve destruction of trackage do not fully interrupt a dispatcher's traffic.*

This view perhaps reflects attitudes and practices pertaining to long-distance, freight-dominated railway operation. Long-distance, low-frequency passenger

services through sparsely-populated areas are not conducive to the use of substitute bus services, as often happens in the UK context (Network Rail, 2004a, p6). While passengers can usually ‘trans-ship themselves’ to other vehicles and modes, if necessary, this is not the case with freight, particularly the heavy, bulky freight for which rail is particularly suited, and so it is imperative that trains continue to move if at all possible (however, Network Rail (ibid) does acknowledge “the need to keep time sensitive goods moving towards their destinations, including by use of ... other modes of transport”). Even if the load carried by a train is not of a time-critical, ‘just-in-time’ nature, keeping trains moving helps to reduce congestion and the build-up of queues and delay.

### **Basic Disruption Management ‘Tools’**

- 5.14 As noted above, even comparatively minor perturbations to heavily-used, frequent passenger train services may result in the overcrowding of trains and/or stations, with implications for passenger comfort, convenience and safety. Such short-term fluctuations in capacity may be dealt with by re-timing trains and/or adjusting their stopping patterns. If the problem results from a train fault, the obvious thing to do is to remove the train from service and/or repair the fault as soon as possible, to remove its effect on other services. A similar approach may be applied to longer-distance main line passenger and freight services, to reduce the impact of a malfunctioning set of rolling stock on the timetable as a whole. A faulty passenger train may skip some or all of its remaining scheduled stops in order to remove it, and its effects on other services, from the system as early as possible. The inconvenience to directly-affected passengers or freight customers may (indeed, should, in such cases) be outweighed by the benefits to other users of the system.
- 5.15 In cases where an infrastructure or total train failure, or some ‘external’ event reduces practical line, route or network capacity to a level less than the prevailing traffic demand, the capacity should obviously be restored as quickly as possible, by removing the obstacle and/or repairing the failure. While the capacity reduction is in place, it is usually desirable to keep as much traffic moving as is safely possible, as advocated by White, quoted in para. 5.13,

above. The approach adopted for dealing with the situation will depend on a range of underlying factors, including the ratio of traffic demand to residual capacity, the operating flexibility of any track(s) remaining in use, the availability of parallel, alternative routes and any restrictions imposed by those alternatives in terms of train types and driver route knowledge (the residual capacity is in fact partly a function of these last three factors).

- 5.16 If the demand is not greater than the practical residual capacity, it should be possible to continue operating the scheduled service pattern in a normal or near-normal fashion, although there will of course be less ‘slack’ available to cope with any further service perturbations or disruptions. If the demand exceeds the practical residual capacity, delays will be incurred, which will increase exponentially as demand approaches the maximum (i.e. theoretical) residual capacity, and queues of trains will form. If demand exceeds the maximum residual capacity, the queues of trains will continue to increase in length until the capacity is partly or wholly restored, or the demand is reduced. As noted in Chapter 2, the effects of such incidents can spread rapidly and widely, possibly affecting trains that do not use the directly affected section of the network if trains queue across junctions.
- 5.17 If the reduction in practical capacity is short-term and to a level not very much less than demand, it may be possible to operate all the scheduled train services on the affected section of the system, albeit with some delay. In this case, consideration will need to be given to the order in which trains pass through the resulting ‘bottleneck’, to ensure that no services are unduly delayed. Otherwise, it will be necessary to reduce demand to a level closer to the remaining practical capacity. This may be achieved by one or more of the following means:
- Diverting trains via one or more parallel routes.
  - Turning services back before they reach the bottleneck (referred to in the British operating context as ‘spinning’).
  - Cancelling services.

- 5.18 The first option obviously depends on the availability of parallel routes, and also of spare capacity on routes that are available. The option may be restricted by limitations imposed by driver route knowledge and the suitability of the route(s) for the trains that require diversion (most obviously, electric trains cannot use routes without the appropriate form of electrification, unless they are hauled by a route-compatible train or locomotive). The situation may also be complicated by a need for trains (particularly passenger services) to call at and/or provide connections at a station or other point on the original route that is not served by the diversion. The diversion option may be particularly suitable for long-distance freight services (Network Rail, 2004a, p6) without intermediate stops.
- 5.19 The second option can provide a passenger service over much of the route, with onward connections being provided by residual through services and/or a replacement/supplementary bus service. Reversing trains on either side of the bottleneck may use some additional capacity on those sections of the route, but that should not be a problem, as long as the remaining capacity is not less than that in the adjacent bottleneck, in which case an extended bottleneck would be created. The third option, cancellation, is the most drastic of the three, but perhaps the simplest to implement. If passenger services are frequent and only some are cancelled, and/or there are route or mode alternatives available, users may not be unduly inconvenienced. Neither the second nor the third option is suitable for freight services, and should be avoided where possible.

### **Disruption Management on Britain's Mainline Railways**

- 5.20 In the context of heavy rail operations in Britain, train movements are regulated by Network Rail's signalling staff. Network Rail's Train Regulation Policy (Network Rail, 2004a, p5) states that "train regulation is a key part of the signaller's role [and] a necessary activity when trains are running outside the planned schedule and are causing pathing conflicts." The Policy goes on to describe the objective of train regulation, as set out in the train operators' Track Access Conditions, as

*striking a fair and reasonable balance between:*

- *minimising overall delay to train movements;*
- *minimising overall delay to passengers and time sensitive goods;*
- *maintaining connections between railway passenger services;*
- *avoiding undue discrimination;*
- *protecting the commercial interests of Network Rail and each affected train operator; and*
- *the interests of safety and security.*

These are all worthy objectives (although the fourth and fifth are perhaps somewhat nebulous), and they closely reflect the train regulation objective established by the Office of Rail Regulation (2004b, p.H16). However, given that passenger trains are of varying capacity and are likely to be unevenly loaded, the first and second objectives may be mutually exclusive. Similarly, maintaining connections is likely to increase overall train, and possibly passenger, delay, since it inevitably requires trains to be held to meet late-running services. It is also difficult to make trade-offs between delays to passengers and time-sensitive goods, without knowing the respective values (costs) of those delays (all goods are time-sensitive to some degree, both in themselves and in terms of the costs associated with the rolling stock and infrastructure being used for their conveyance).

5.21 It can be seen from the preceding paragraph that regulation is performed with the objective of minimising delay to trains, to passengers and to freight. Taking account of overall passenger and freight delay entails consideration of the relative loadings and priorities of different types of trains, as well as the extent of total and individual train delay. Information about freight train loadings should be quite readily available prior to and during their journeys, but real time passenger loading data is more problematic; it may be possible to obtain approximate such data from ticket sales and automatic barrier data, and, as in the case of London Underground operations, from measurements of train loadings

in terms of total passenger weights. For relative train priorities, the set of train classes currently used in British railway operations is listed in Table 5.1, below.

**Table 5.1: Train Classifications in British Railway Operations**

<b>Train Class</b>	<b>Train Types in Class</b>
1	Express passenger train; Nominated postal or parcels train; Breakdown or overhead line equipment train going to clear the line or returning therefrom (1Z99); Traction unit going to assist a disabled train (1Z99); Snowplough going to clear the line (1Z99); Class 9 - 373/1 or 373/2 train [Eurostar and Regional Eurostar]
2	Ordinary passenger train; Breakdown or overhead line equipment train not going to clear the line (2Z99); Officer's Special train (2Z01)
3	Freight train capable of running at more than 75mph (121km/h), or a parcels train or an empty coaching stock train where specially authorised
4	Freight train permitted to run at more than 60mph (97km/h)
5	Empty coaching stock train
6	Freight train permitted to run at 50, 55 or 60mph (81, 89 or 97km/h)
7	Freight train permitted to run at 40 or 45mph (65 or 73km/h)
8	Freight train permitted or timed to run at 35mph (57km/h) or less
9	See Class 1. Class 9 formerly comprised non-fully fitted/unfitted/non-continuously braked freight trains
0	Light locomotive(s)

(Sources: Hall, 2001, p37; National Economic Research Associates and Symonds Group, 2000, p12; Lu, 1996-98)

This classification dates back to the early 1960s, when four-character train descriptions, or headcodes, were introduced by British Rail, with the first digit in a train's headcode indicating its Class (Simmons and Biddle, 1997, p202). The classification reflects the much greater variety of freight that was then conveyed by rail than is now the case. It can be seen that the classification represents a fairly obvious hierarchy, with train Classes declining in priority from 1 to 9 and then 0. Network Rail staff have indicated that only Classes 1 and 2 are now explicitly used for regulation purposes, with Class 1 trains normally being given priority over their class 2 counterparts when a routing

conflict occurs. This is not ‘written in stone’, however, and Class 2 services may on occasion be given priority over their Class 1 counterparts (see below).

5.22 Network Rail’s existing regulation policy, dating from 1996, comprises three levels of policy (Network Rail, 2004a, p5):

- *level 1: generic instructions applicable at any signalling location, which is the default policy;*
- *level 2: instructions for generic groups of train services applicable at a particular location where trains may be regulated; and*
- *level 3: instructions for specified trains (by unique identifier) at particular locations where trains may be regulated.*

Discussions with Network Rail staff have indicated that the level 1 policy is to minimise overall train delay, the first of the train regulation objectives listed above in para. 5.20.

5.23 According to Network Rail (2004a, p6),

*In practice, the application of the [regulation] policy has become more challenging with the increased range of speed and acceleration characteristics of new trains, and with changes in the access rights of separate train operators and there is widespread industry recognition of the need to alter policy to take account of these factors and a need to simplify and aid the signaller’s task.*

In response to these issues, it has been decided to develop and introduce “a revised regulation policy, [comprising] three key components”:



- *a generic regulation statement, applicable across the network, that defines the regulation objective and the criteria to be applied by signallers in arriving at regulation decisions;*
- *the introduction of margin tables as a key information aid to signallers that readily identifies the relative performance characteristics of particular traction types and available infrastructure for regulation purposes; and*
- *the introduction of an industry-agreed protocol for the management of long-distance trains on the network. Such a protocol will recognise that long-distance trains are vulnerable to a series of uncoordinated localised regulating decisions that have a tendency to transmit reactionary delays across the network.*

Again, it is understood that the generic regulation statement will define the regulation objective as being to minimise overall train delay.

5.24 As a partial aside, Wolmar (2001, pp134-135) has the following to say about train regulation in the post-British Rail operating environment:

*Another decision resulting from the break-up of the railways ... was the change ... in the priority given to different types of trains. Under BR, high-speed passenger trains always had priority over goods trains, which sometimes sat in sidings for a long time before being allowed to proceed. [That was] changed ... after consultation with the industry, creating a system of minimum overall delay. ... Railtrack, keen to minimise the cost of delays, ran 'track access awareness' sessions for signallers, which outlined the penalty system. The new rule ... said that various factors had to be balanced in train regulation, including 'protecting the commercial interests of Railtrack and each affected train operator'.*

Wolmar goes on to say that this requirement was removed in November 2000, apparently as a consequence of safety concerns which were highlighted by the Southall crash in September 1997, where a freight train was routed across the path of a high-speed passenger train (while there was nothing inherently wrong with the freight train routeing, the consequences of the passenger train passing a signal at danger were very much worse than would otherwise have been the case). However, the current regulation objective quoted above in para. 5.20 appears to contradict this.

- 5.25 From the preceding paragraphs, it can be seen that perturbed train movements are regulated with the objective of minimising overall delay to trains, passengers and time-sensitive (i.e. all) freight, while simultaneously seeking to maintain connections, protect the affected commercial interests in an even-handed fashion and maintain standards of safety and security. Network Rail is introducing network control centres where Network Rail and Train Operating Company (TOC) staff are co-located (Network Rail, 2004a, p8). This approach helps to ensure the protection of the commercial interests of the affected companies when responding to perturbations, particularly in situations where trains have to be ‘spun’ or cancelled, and accelerates the decision-making process. Where two or more TOCs may be affected by regulation decisions, as in the area controlled from Waterloo, for example, an ombudsman may be employed to protect the interests of the non-dominant TOCs (in Waterloo, South West Trains is the dominant TOC, but the section of the network controlled from there is also used by passenger trains operated by Virgin CrossCountry, Southern and Eurostar, and by freight trains operated by EWS and Freightliner).
- 5.26 In the course of an informal visit to the Waterloo control centre in July 2004, the author was advised that train regulation is done on a pragmatic basis, rather than by means of hard-and-fast rules. Where possible, agreement is reached between the interested parties as to what should be done in a given situation. In situations where capacity is reduced, the effects are shared as equably as possible among the affected operators, making due allowance for freight requirements. Informal methods are used to reduce overall passenger delay and inconvenience – a lightly-loaded Class 1 service might be terminated prior to

reaching its destination, and the passengers transferred onto a higher-capacity Class 2 service, since it makes little sense to do the opposite. For example, an Exeter – Waterloo service might be terminated and reversed at Basingstoke, with the passengers transferring to a service from Southampton or Portsmouth.

- 5.27 This approach seems sensible, and well-suited to situations where major disruptions occur, seriously amended, degraded or emergency working conditions apply (see Table 2.2 for definitions), and strategic decisions have to be made about diverting, spinning and/or cancelling services. It is doubtful whether a single best response can easily be identified in such situations, and it may be that previously-arranged area-specific contingency plans (Network Rail, 2004a, p7), modified as necessary, are the most appropriate means of dealing with such situations. These are analogous to ‘integrity envelopes’ used in the offshore industry (Lloyd’s Rail Register, 2004, p1) to

*ensure that potential failures and possible responses, including alternative systems and methods of working, are identified in advance so that if a failure happens, staff know what the options are, what they are to base their decisions on, and how long they should operate in that state.*

At the other end of the ‘perturbation scale’, an individual signaller or controller should be able to make the appropriate regulatory decisions in response to a minor disruption of services. Between these two extremes it seems likely that there is a ‘middle ground’ of amended and degraded working conditions, where services can be maintained without the need for cancellations or the spinning of trains, but where the situation and number of affected trains is such that an individual (or individuals) cannot reliably make the optimal regulatory decisions, but where the implementation of contingency plans and ‘regulation by committee’ is not an option, either. In such situations, the provision of some sort of computerised regulation ‘assistance’ is likely to be of considerable benefit.

- 5.28 It appears that the use of margin tables to implement the revised regulation policy will place the emphasis on reducing delay, rather than prioritising movements by train class, and that train priorities will be addressed by the third policy component, which will make allowance for the needs of Class 1 services and other long-distance freight trains. This is consistent with the example given of the approach adopted in the Waterloo control centre. Such a hybrid approach, taking account of train delays and relative priorities, is needed if both overall delay and individual passenger/freight delays are to be minimised, as noted in para. 5.21.
- 5.29 White (2003b) makes the following observations about the roles of delay (pp22, 56) and priority (p56):

*Delay measurement is used to assess the efficiency of operation and the capacity of the infrastructure. Delay is also generally the factor used to determine how trains will interact with each other during planning and execution of the plan [cf Pachel, quoted in para. 5.6].*

*The definition of delay that is used in evaluating train operation will affect the planning process significantly. If a train that leaves and/or arrives on time at schedule points is not considered to have been delayed regardless of slowing or stopping between schedule points, planning and executing the plan is relatively straightforward. If trains are considered delayed if they do not make minimum possible running time regardless of schedule, planning and executing the plan can become very difficult when train movement instructions prohibit delay to certain trains. Instead of being one of the components of a plan, delay becomes the only component of a plan. The other elements of planning become obstacles that constantly interfere with the plan because they were not considered when the schedule was designed.*

*Priority is a very important concept in traffic planning and in handling trains. Priority is most effective when designed into the timetable, not when it is the determining factor in operation regardless of the schedules. Depending upon how priority is used on a given railroad, it may be a tool or the single defining element of planning and operation. Among the elements of planning, the implementation of delay and priority have the greatest effect. When priority is the greatest consideration and delay has the most strict definition, the result can easily be congestion that causes extreme delay to all but a few important trains. The congestion caused by basing all decisions on priority can result in unavoidable delays to high priority trains.*

The only way a train can remain on time at its scheduled timing points without minimising its intermediate running time(s) is to build some slack, or ‘recovery time’ into the timetable. While this provides operating flexibility, it also consumes capacity. White’s comments reinforce the complementary roles of train delay and train priority in the regulation of train movements, and the importance of not over-emphasising the significance of one at the expense of the other.

- 5.30 As noted in Chapter 2 and in para. 5.11 above, the SRA has suggested that prioritisation rules and class regulation practices should be revisited, with the aim of improving train regulation. The existing classification system for trains using the British railway network (see para. 5.21 and Table 5.1) is somewhat outdated, and of relatively little use for train regulation purposes, as implied by the first quotation in para. 5.23 and illustrated by the fact that only two of the classes appear to be used in practice (this is compounded by the fact that some TOCs have apparently been seeking the re-classification of some Class 2 trains as Class 1 services). The classification system adopted also has some serious fundamental shortcomings, in that it uses an ordinal rather than interval/ratio scale, which is of limited use in comparing train priorities. The difference between the two scaling systems is summarised by De Vaus (1996, pp130-131) as follows:

*an ordinal variable is one where it is meaningful to rank the categories: there is some justifiable order between the categories. However, it is not possible to quantify precisely how much difference there is between the categories. ... Any variable in which categories can be ranked but where the difference between the categories cannot be quantified in precise numerical terms is an ordinal variable.*

*An interval/ratio variable is one in which the categories have a natural ranking and it is possible to quantify precisely the differences between the categories. Age, if it is measured in years, is an interval variable because as well as ranking people according to their age, the precise difference between the ages can be quantified. If age was simply measured as young, middle-aged and old it would only be an ordinal variable.*

If two trains of different classes are competing for a path, or ‘slot’, on a route, choosing between them is straightforward, all other things being equal.

However, if a class 2 train is competing with a succession of class 1 trains, it does not seem appropriate that it should have to cede priority to each class 1 train in turn, as strict prioritisation by train class would imply. Similarly, it is impossible to determine objectively whether a 2-minute delay to each of two class 2 trains is equivalent to a 2-minute delay to a single class 1 train, or, if not, which is preferable. In essence, as De Vaus indicates, it is impossible to determine how much more or less important one class of train is than another.

- 5.31 If two trains of the same class are competing for a single path, there is again no obvious means of choosing between them on the basis of class priority. However, two trains should not be timetabled to be using the same section of the network at the same time; it follows that, in the circumstances, and barring a timetabling error, at least one of the trains is running ‘out of course’, i.e. running either ahead of or behind schedule. In such cases, one solution would be to give precedence to the train that is experiencing the most delay (if one train is on schedule and the other is early, the early train may be said to have accumulated

‘negative delay’, in which case the on-time service should proceed first).

However, if the ‘less late’ train is not running early, it is likely to be delayed as a result, so that an on-time service will incur a delay, or an already-late service will be delayed further.

5.32 The situation will be exacerbated if a slower-running train is routed ahead of a faster service, since the faster service will accumulate further delay while running behind the slower one. The use of margin tables should help to avoid this situation, since faster-running services (typically of equal or higher classification to slower-running trains) will normally be given priority. This is analogous to the ‘Shortest Processing Time first’ (SPT) rule in scheduling theory (French, 1982, pp35-36; Pinedo, 2002, p59), and should have the effect of minimising the overall delay to train movements, in accordance with Network Rail’s (and the ORR’s) objectives (see para. 5.20, above). There may be situations where a regulation decision based on train classification contradicts the equivalent decision based on the use of margin tables; in such cases, Network Rail’s revised regulation policy (see para. 5.23) suggests that the margin table-based regulation decision would be used.

5.33 There remain two problems with the margin table approach to regulation:

- (i) Routing a lightly-loaded but faster service ahead of a heavily-used slower service will inevitably increase the delay to the second train. Although this will minimise (for the two trains in question) the overall delay to train movements, it may increase the overall passenger delay, contrary to the second objective stated in para. 5.20 (it is assumed that these hypothetical trains are passenger-carrying, but similar considerations apply to freight).
- (ii) Where trains are queuing on each of two converging routes to enter a single section of track and there are more than two trains in total (i.e. there are at least two trains waiting on at least one of the competing upstream routes, although one or more of the trains in the rear might be approaching a signal set to Danger, rather than being stationary), the

‘logical’ choice between the first two trains may not give the best overall result. As noted in para. 5.30, regulating strictly by train class in this situation could produce some perverse results, and ‘pair-wise’ regulation by margin table should at least minimise the train delay at each step of the regulation process, if not overall. However, as in (i), it has the potential to increase overall passenger delay in situations where lightly-used, faster services are routed ahead of slower-moving, heavily-used trains.

It can thus be seen that existing methods of regulation have some significant shortcomings. Possible means of addressing these are considered in the next section.

### **Proposed Framework for Disruption Management**

5.34 The foregoing material indicates that there are three major issues relating to the regulation of trains in Britain:

- (i) The objectives of minimising overall train delay and minimising overall passenger/freight delay may be mutually exclusive, i.e. regulating trains by means of margin tables to reduce overall train delay may result in an increase in overall passenger/freight delay;
- (ii) Regulating trains on the basis of their priorities is based on very restricted and ‘coarse’ train classifications (which additionally suggest that all freight trains other than nominated postal or parcel trains are of a lower priority than any passenger train), and may again result in an increase in overall train delay and/or overall passenger/freight delay;
- (iii) Regulating multiple (i.e. more than two) trains by either means, using a series of comparisons between successive pairs of trains, will not necessarily yield an optimal result.

In response to these issues, an alternative approach is proposed, as set out below.





- 5.35 The first two issues could be dealt with by assigning an appropriate weighting to each train using the system, reflecting its social value and thus the social cost of any delay to the service. Such an approach would provide a finely-graded train classification, covering both passenger and freight services, and would reflect the SRA's observation (2003, p11) that "in practice, the costs and the benefits of each train will differ." A revised regulation objective of minimising overall weighted train delay would then reconcile the sometimes conflicting aspirations of minimising overall train delay and also minimising overall passenger delay. Weightings are already employed in ARS (Goodman and Takagi, 2004, p772).
- 5.36 The third issue could be dealt with by means of classical optimisation techniques such as 'hill-climbing', whereby train sequences would be successively adjusted with the objective of the minimising the objective function of overall weighted train delay, subject to various constraints (including, for example, maintenance of connections between services, whereby the relevant departures could not occur until their connecting services had arrived and the minimum connection time for passengers had elapsed). However, it is proposed here instead to use recognised specialist sequencing and scheduling techniques, whose theory is typically based on industrial production applications, but can also be applied to transportation issues. In the words of Conway et al (1967, p1):

*Sequencing problems are very common occurrences. They exist whenever there is a choice to the order in which a number of tasks can be performed. A problem could involve: jobs in a manufacturing plant, aircraft waiting for landing clearances, bank customers at a row of tellers' windows, programs to be run at a computing center, or just Saturday afternoon chores at home. Our basic thesis is that, regardless of the character of the particular tasks to be ordered, there is a fundamental similarity to the problems of sequence.*

These proposals are now developed in more detail.

## Regulation to Minimise Overall Weighted Train Delay

5.37 The first two regulation issues identified in para. 5.34, and the solution proposed in para. 5.35, are illustrated below, by means of a hypothetical example.

**Figure 5.5: Train Regulation Example 1**

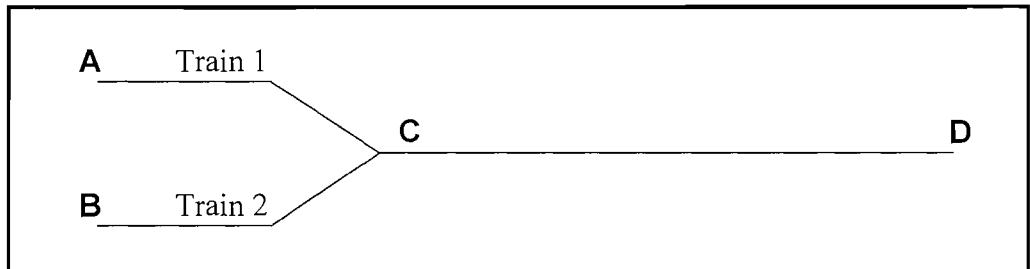


Figure 5.5 shows two trains, numbered 1 and 2, which are travelling from A and B, respectively, to D via C. The section of line between C and D can only be occupied by one train at a time. For the purposes of simplicity, it is assumed that neither train experiences any delay until one of them occupies the section of route between C and D (this is equivalent to ignoring any delay that has occurred until this point, and then seeking the approach that minimises the subsequent delay). It is further assumed that Train 1 is carrying 120 passengers and takes 12 minutes to travel from C to D, while Train 2 is carrying 45 passengers and takes 9 minutes to travel from C to D. It is also assumed that Train 1 is of Class 2, and that Train 2 is of Class 1, reflecting their respective speeds. These assumptions are summarised below in Table 5.2.

**Table 5.2: Train Data**

Train No.	Train Class	No. of Passengers	Travel Time (mins)
1	2	120	12
2	1	45	9

5.38 If Train 1 proceeds first, Train 2 experiences train delay of 12 minutes, and passenger delay of 540 minutes. If Train 2 proceeds first, Train 1 experiences train delay of 9 minutes, and passenger delay of 1080 minutes. Regulating by train class minimises train delay, but maximises passenger delay, as does regulating by means of the margin table approach, as applied to train travel times. Regulating to minimise overall passenger delay, on the other hand,

maximises overall train delay. All three approaches thus produce contradictory indications in terms of Network Rail’s regulation objectives of minimising overall train delay and minimising overall passenger delay.

5.39 It is proposed to reconcile these differences by applying a suitable weighting to each train. Such a weighting should reflect the passenger or freight loading of each train, and the comparative ‘importance’ of each train, as expressed by its classification, for which speed is a possible proxy. For purposes of illustration, each train is therefore weighted according to the number of passengers it is carrying, divided by its journey time between C and D. The resulting weighted train delay values for Trains 1 and 2 are then 90 and 60 minutes, respectively, and the weighted delay is minimised by allowing Train 1 to proceed ahead of Train 2. The results are summarised in Table 5.3, with the minimum train, passenger and weighted delay values highlighted in bold. This clearly illustrates the potential conflict between the objectives of minimising train delay and minimising passenger delay. As already noted, the scenario and the values used are completely hypothetical, but serve to illustrate the general point.

**Table 5.3: Regulation Outcomes**

Train Sequence	Delay (minutes)		
	Train	Passenger	Weighted
1, 2	12	<b>540</b>	<b>60</b>
2, 1	<b>9</b>	1080	90

5.40 As acknowledged by Nash et al (2004), establishing the value and appropriate costs of train paths is a complex process, particularly as the appropriate valuation varies by location and time. Establishing the costs associated with delays to different trains is therefore potentially difficult, although values of time for different categories of traveller have been established for the purposes of the economic assessment of transport investment proposals, and it seems likely that the costs of delays to freight have been determined within the logistics industry.

5.41 Under the terms of their Track Access Agreements, Train and Freight Operating Companies are compensated for delays to their services caused by Network Rail

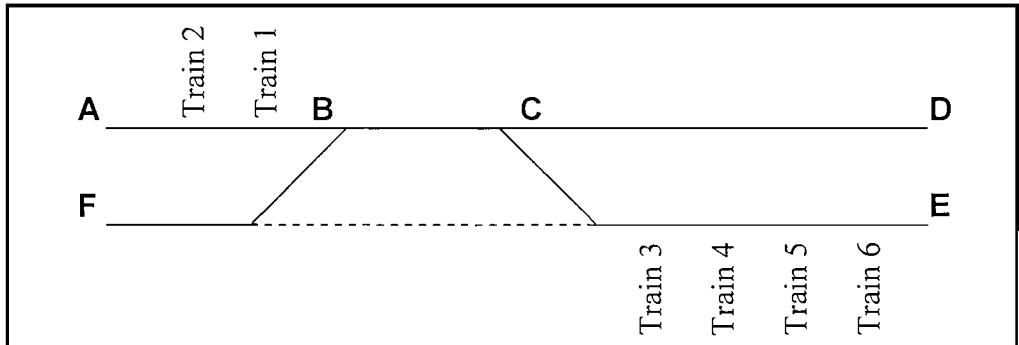
or other Operating Companies, and it is understood from conversations with Network Rail staff that these penalty payments per delay minute vary between different trains. The values of these payments appear to be confidential, but would seem to be appropriate as weightings for delays to the respective trains, the total weighted value of which could then be targeted for minimisation. The same conversations have indicated that Network Rail does not want to, or feels it cannot, be seen to regulate so as to minimise its penalty payments, and thus to maximise its own interests. However, if the trains using the system are assigned the appropriate delay minute costs, regulating in this manner should be in society's (as well as Network Rail's) best interest.

#### Regulating Larger Numbers of Trains (More than Two)

- 5.42 Despite the issues raised above, making a regulatory decision in the context of two competing trains is a fairly trivial exercise: if the various criteria for regulation fail to produce a clear outcome, an arbitrary choice can be made, probably (given that there is no clear distinction between them) without causing undue delay to either train. However, when more than two trains are under consideration, the regulation task increases in complexity.
- 5.43 It has already been observed (see paragraphs 5.29 and 5.30) that the use of strict priority rules could cause a large amount of delay to a train of low classification. In addition to the likelihood of causing undue delay to such a train, there is the possibility that one or more trains of higher classification might be waiting in its rear, the delay to which would be of still greater concern, but would not be considered by a simple pair-wise comparison between the train in advance and successive trains on a converging route. A similar consideration applies to successive pair-wise comparisons between trains on converging routes when using the margin table approach to minimise overall train delay. This is illustrated by another hypothetical example, illustrated in Figure 5.6, below.
- 5.44 The Figure shows a two-track railway, with lines running from A to D and E to F. A section of the line from E to F has been blocked (indicated by the dashed line), forcing trains to be re-routed via the crossovers to and from C and B. The result of this is that trains in both directions have to use a single, shared section

of track between B and C. This scenario could also occur as the result of a conflict at a bottleneck on a route, such as a single-lead junction, caused by poor timetabling or trains running out of order.

**Figure 5.6: Train Regulation Example 2**



5.45 In this example, two trains, numbered 1 and 2, are waiting at the approach to the ‘B’ end of the temporarily-shared section of track, travelling from A to D, and four trains, numbered 3 to 6, are waiting at the approach to the ‘C’ end, travelling from E to F. The times required by each train to enter, traverse and clear the shared section of track are shown in Table 5.4. These times are again completely hypothetical, and have been chosen to illustrate the point made in paragraphs 5.42 and 5.43 about multi-train regulation by means of margin tables.

**Table 5.4: Train Travel Times**

Train Number	Travel Time (minutes)
1	6
2	3
3 - 6	5

In reality, the queuing trains would normally be separated by signals at Danger, with each train occupying a single block. As the first train in either queue entered the shared section of track and cleared the signal overlap, the train(s) in its rear would move forward in succession. For the purposes of simplification, it is assumed that when a train clears the shared section of track (or the overlap of the signal in advance of the shared section), the train in its rear is in position to enter the shared section of track if given a signal to do so. In reality, this might

not be the case. As in the previous example, delay is calculated from the time at which the first train enters the shared section of track (or, more accurately, passes the signal protecting the shared section).

5.46 The first decision to be made is whether to allow Train 1 or Train 3 to enter the shared section first. Regulating by margin table (or by classification, assuming a train's status in the classification hierarchy to reflect its speed, and thus to be inversely related to its travel time), Train 3 is routed ahead of Train 1, and the process is repeated until Train 6 has cleared the shared section, at which point Train 1 and then Train 2 resume their journeys towards D (of course, this assumes that no other, higher-classified, trains arrive from E in the meantime). Regulating strictly by class or by train travel time, this procedure is correct, but it ignores the fact that four F-bound trains are being routed ahead of a faster/higher-classified train (Train 2) as well as the slower/lower-classified Train 1. The consequences of this, in terms of approximate total train delay, are described in the following paragraphs.

5.47 In order to estimate the delays incurred by individual trains and in total, the following procedure was adopted: while a train is travelling through the shared section, any trains waiting to travel in the opposite direction are assumed to incur delay equivalent to the travel time of the moving train. Any trains in rear of the moving train are assumed to incur delay equal to the travel time of the train in the shared section, minus their own travel times, since a slower-moving train will typically retard a faster-moving one. If the train in the shared section has a travel time less than or equal to a train in its rear, the train in rear is assumed to incur no delay as it moves forward. The delays incurred as a result of the regulating sequence described in the preceding paragraph are shown in Table 5.5. The rows contain the delays incurred by each train during the movements of the others, which are listed in order of movement across the top of the table. Thus, the rows show the individual and total delays incurred by each train as a result of the others, while the columns list the individual and total delays inflicted by each train on the others.

**Table 5.5: Delay Resulting From Strict Regulation By Margin Table**

Delayed Trains	Delaying Train Numbers, In Order of Movement						Totals
	3	4	5	6	1	2	
1	5	5	5	5	0	-	20
2	5	5	5	5	3	0	23
3	0	-	-	-	-	-	0
4	0	0	-	-	-	-	0
5	0	0	0	-	-	-	0
6	0	0	0	0	-	-	0
<b>Totals</b>	10	10	10	10	3	0	<b>43</b>

(Note ‘-’ in the table indicates that the train in question has previously been ‘processed’, i.e. has completed its journey through the shared section, and thus incurs no further delay.)

It can be seen from the table that the total delay incurred is 43 minutes, and that the train incurring the greatest delay, 23 minutes, is train number 2, which has the highest ‘classification’ of all. It seems likely that this situation could be improved upon.

- 5.48 The sequencing procedure was repeated, routing Train 1 through the shared section first, followed by Train 2 (re-adopting the margin table/classification approach), and then Trains 3 to 6. The results are shown in Table 5.6.

**Table 5.6: Delay Resulting From Re-Sequencing Trains (1)**

Delayed Trains	Delaying Train Numbers, In Order of Movement						Totals
	1	2	3	4	5	6	
1	0	-	-	-	-	-	0
2	3	0	-	-	-	-	3
3	6	3	0	-	-	-	9
4	6	3	0	0	-	-	9
5	6	3	0	0	0	-	9
6	6	3	0	0	0	0	9
<b>Totals</b>	27	12	0	0	0	0	<b>39</b>

It can be seen that this sequence of operations results in a 10% reduction in total delay, and also in a more even distribution of delay, in that the maximum train delay is 9 minutes, compared with 23 in the initial sequence. The procedure was repeated three more times, routing Trains 1 and 2 after Trains 3, 4 and 5, in turn. The results of these repetitions are shown in Tables 5.7, 5.8 and 5.9.

**Table 5.7: Delay Resulting From Re-Sequencing Trains (2)**

Delayed Trains	Delaying Train Numbers, In Order of Movement						Totals
	3	1	2	4	5	6	
1	5	0	-	-	-	-	5
2	5	3	0	-	-	-	8
3	0	-	-	-	-	-	0
4	0	6	3	0	-	-	9
5	0	6	3	0	0	-	9
6	0	6	3	0	0	0	9
<b>Totals</b>	10	21	9	0	0	0	<b>40</b>

**Table 5.8: Delay Resulting From Re-Sequencing Trains (3)**

Delayed Trains	Delaying Train Numbers, In Order of Movement						Totals
	3	4	1	2	5	6	
1	5	5	0	-	-	-	10
2	5	5	3	0	-	-	13
3	0	-	-	-	-	-	0
4	0	0	-	-	-	-	0
5	0	0	6	3	0	-	9
6	0	0	6	3	0	0	9
<b>Totals</b>	10	10	15	6	0	0	<b>41</b>

**Table 5.9: Delay Resulting From Re-Sequencing Trains (4)**

Delayed Trains	Delaying Train Numbers, In Order of Movement						Totals
	3	4	5	1	2	6	
1	5	5	5	0	-	-	15
2	5	5	5	3	0	-	18
3	0	-	-	-	-	-	0
4	0	0	-	-	-	-	0
5	0	0	0	-	-	-	0
6	0	0	0	6	3	0	9
<b>Totals</b>	10	10	10	9	3	0	<b>42</b>

5.49 It can thus be seen that strict adherence to the use of margin tables and/or train classification does not necessarily lead to the minimisation of train delay, much less passenger delay, in multi-train regulation scenarios. It can reasonably be inferred that the application of weightings to trains will not necessarily improve this situation, since an equivalent situation could easily arise, whereby a train with a greater weighting was ‘masked’ by one with a lesser one. The type of calculations performed here, even in the course of assessing a relatively straightforward situation, with a small number of trains and various simplifying



assumptions, are too time-consuming to be done manually in practical applications, and the permutations involved are almost certainly too complex to enable the preparation of multi-train margin tables. However, the situation lends itself to computer analysis, if the relevant data are readily available to the necessary software and up-to-date.

- 5.50 Reference has already been made (para. 5.36) to the relevance of existing sequencing and scheduling techniques to these issues. Pinedo (2002, p1) describes the role of scheduling thus:

*Scheduling deals with the allocation of scarce resources to tasks over time. It is a decision-making process with the goal of optimizing one or more objectives.*

*The resources and tasks in an organization can take many forms. The resources may be machines in a workshop, runways at an airport, crews at a construction site, processing units in a computing environment and so on. The tasks may be operations in a production process, take-offs and landings at an airport, stages in a construction project, executions of computer programs, and so on. Each task may have a certain priority level, an earliest possible starting time, and a due date. The objectives can take many forms. One objective may be the minimization of the completion time of the last task, and another may be the minimization of the number of tasks completed after their respective due dates.*

In the present context, the scarce resource is track capacity, and the tasks to be performed are the routing of trains along that track, with various objectives, as set out in para. 5.20, chief among which are the minimisation of delay to trains and/or passengers and/or freight.

5.51 Pinedo (op cit, p14) gives the following definitions:

- *Release date. The time a job arrives at the system (i.e. the earliest time at which [it] can start its processing).*
- *Due date. The committed shipping or completion date (the date the job is promised to the customer).*
- *Weight. The weight ... is basically a priority factor, denoting the importance of [a] job ... relative to other jobs in the system.*

For each train, its release date is the time at which it reaches its point of access to the section of track whose capacity is scarce; its due date is the time at which it is scheduled to leave that section of track; and its weight is its classification, which, as implied by the definition, must take the form of a weighting as described above in para. 5.30, i.e. one measured on an interval/ratio scale.

5.52 French (1982, pp9-10) describes some of the objectives of the scheduling process in general terms, of which those most relevant to the present context are:

- *Keep promised delivery dates [i.e. due dates, or the times at which the trains are scheduled to leave the section of the network under consideration].*
- *Minimise the overall length of the scheduling period [so that the resources] may be released for other tasks [this also has the obvious benefit of minimising the time required to move late-running trains through the congested section of the network].*
- *Minimise the time for which the machines [i.e. track, in the present context] are idle; idle machines mean idle capital investment [this should not be an issue when dealing with disruption, but does reflect one of the principles of railway*

*operation identified in paragraphs 2.27 and 2.28 of Chapter 2].*

French (ibid, pp10-11) also provides a list of definitions, of which those most relevant to the present context are:

- *Due date [again, the time at which a train is scheduled to leave the section of the network under consideration].*
- *Completion time ..., i.e. the time at which the processing of [a job  $\equiv$  a train] finishes [i.e. the time at which a train does leave the section of the network under consideration].*
- *Flow time ... is defined to be the time that [a job] spends in the [processing environment. Thus Flow time = Completion time – Release date].*
- *Lateness [ i.e.] the difference between [a job's] completion time and its due date. ... Note that when a job is early, i.e. when it completes before its due date, [its lateness] is negative. It is often more use to have a variable which, unlike lateness, only takes non-zero values when a job is **tardy**, i.e. when it completes after its due date. Hence we also define the tardiness and, to be comprehensive, the earliness of a job.*
- *Tardiness [a job's lateness or 0, whichever is greater].*
- *Earliness [a job's lateness, negated, or 0, whichever is greater].*

5.53 Using these definitions, Pinedo (2002, pp18-19) lists examples of possible objective functions for minimisation, of which the relevant ones are:

- *Makespan. The makespan ... is equivalent to the completion time of the last job to leave the system. A minimum*

*makespan usually implies a high utilisation of the machine(s).*

- *Maximum Lateness. [This] measures the worst violation of the due dates.*
- *Total weighted tardiness. This is ... a ... general cost function [which is equivalent to the total weighted delay experienced by trains, the value we are seeking to minimise].*

The total weighted tardiness is equivalent to the total weighted delay experienced by trains, the value we are seeking to minimise. The use of the tardiness, rather than the lateness, value is appropriate, since there is no benefit to be had from trains running ahead of their schedules (in any case, in the situations being dealt with in the present context, it is unlikely that any train would be able to do so).

- 5.54 For ‘Single Machine Processing’ situations (equivalent to the situations under consideration here, where trains on two either normally or temporarily converging routes have to be ‘processed’ by a single, shared section of track), two simple approaches to developing a schedule are Shortest Processing Time (SPT) scheduling and Earliest Due Date (EDD) scheduling. In the former, jobs are processed in order of increasing processing time (equivalent to the use of margin tables), and in the latter, jobs are processed in the order of their due dates, i.e. the times at which they are scheduled for completion (equivalent to the approach suggested in para. 5.31, where the later-running of two trains goes first). The former approach minimises the mean flow time, and thus also the average completion time and lateness (French, pp38-39). The latter approach minimises the maximum lateness, and thus the maximum tardiness, resulting from the schedule. However, both approaches assume that all jobs are available for processing from the start, and can thus be processed in any order. In the case of each of the two converging queues of trains in the second example above, only the first train in each queue is initially available, and so this assumption does not hold. This is referred to as a precedence constraint, and the jobs thus

affected are said to form a string or a chain (French, 1982, pp48-52; Pinedo, 2002, p35).

- 5.55 Pinedo (2002, p38) describes an algorithm that “minimizes the total weighted completion time when the precedence constraints take the form of chains.” This makes use of the ‘ $\rho$ -factor’ in each chain, which is obtained by calculating for each successive job in the chain the cumulative value of the sum of the weights of that and the preceding jobs in the chain, divided by the sum of their processing times, and finding the maximum cumulative value. For a chain of  $k$  jobs,

$$\rho - \text{factor} = \max_{1 \leq l \leq k} \left( \frac{\sum_{j=1}^l w_j}{\sum_{j=1}^l p_j} \right)$$

where

$w_j$  = the weight of job  $j$ ; and

$p_j$  = the processing time for job  $j$ .

Minimising the total completion time is obviously equivalent to minimising the average completion time of the jobs comprising the schedule, which, in turn, is equivalent to minimising their average flow times and lateness values (French, 1982, p28). The algorithm is as follows (Pinedo, 2002, p38):

*Whenever the machine is freed, select among the remaining chains the one with the highest  $\rho$ -factor. Process this chain without interruption up to and including the job that determines its  $\rho$ -factor.*

This algorithm is now applied to the example described in paragraphs 5.43 – 5.49.

- 5.56 The two chains of jobs in the example are

1 → 2

and

$$3 \rightarrow 4 \rightarrow 5 \rightarrow 6$$

The weights (all = 1 in this case) and processing times of the jobs are shown in Table 5.10.

**Table 5.10: Job Weights and Processing Times**

Job No.	1	2	3	4	5	6
$w_j$	1	1	1	1	1	1
$p_j$	6	3	5	5	5	5

For both chains, the  $\rho$ -factors are determined by the final job in the chain: the values are  $2/9$  and  $4/20$  respectively, equivalent to 0.22 and 0.20. The algorithm therefore indicates that the job sequence should be

$$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6$$

This confirms the result obtained above, and the algorithm provides a general methodology for dealing with such situations, including cases where the trains involved have different weightings, thus addressing the third regulation issue identified in para. 5.34.

- 5.57 This example is comparatively straightforward, and simplifying assumptions have been made about the ‘processing times’ of successive trains (see para. 5.45). However, more realistic times can be calculated reasonably easily and quickly, based on the respective block lengths and train performance characteristics. As long as the congestion persists, the situation is likely to change as more trains approach the bottleneck on the affected routes. The calculated schedule would therefore require successive updating, and, possibly, complete amendment; the latter requirement might arise if a train with a large weighting and small processing time joined a queue and established a new maximum  $\rho$ -factor. The appropriate point at which to include trains in the process is probably when they first start experiencing delay as a result of the congestion ahead. Given that the time taken for each train to clear the shared

section of track is likely to be measured in minutes, particularly where trains have to start from standstill, it should be possible easily to update the schedule as necessary between successive train movements through the bottleneck. Where parallel routes are available, but capacity is still less than demand, ‘parallel machine’ scheduling techniques (Pinedo, 2003, Chapter 5) are available.

- 5.58 The proposed framework helps to meet the first two elements of the train regulation objective quoted in para. 5.20, and to address their potential mutual exclusivity. It also (at least partially) revisits prioritisation rules and class regulation practices, as suggested by the SRA (see para. 5.11, above). There remains scope for a great deal of further work in this area, some of which is outlined below.

### **Further Work**

- 5.59 In terms of the proposed regulatory framework, there is a need for the agreement and establishment of suitable weightings for passenger and freight trains. While these must obviously vary between trains if they are to be of any use, it may well be that the weighting applied to an individual train (e.g. a long-distance passenger service) would vary by time and location along its route, as it competes for scarce capacity with different types of train and with peak and off-peak services.
- 5.60 There may be scope for extending the techniques associated with the regulation of metro-type train services (see para. 5.12) to frequent, intensively used main line services on inner-suburban or longer-distance commuter routes, although the typical comparative complexity of main line operations presents difficulties in this regard (Goodman and Takagi, 2004, p766).
- 5.61 As described in the preceding section, there is scope to extend the application of scheduling techniques to the regulation of trains on converging routes. It may also be possible to apply such techniques to the maintenance of connections between trains (Goodman and Takagi, 2004, p771), the avoidance of undue discrimination between operators, and the protection of commercial interests,

the importance of which was noted in para. 5.20. For example, techniques exist for the minimisation of average flow time, subject to thresholds of maximum tardiness for individual jobs (French, 1982, pp58-63).

- 5.62 As noted in para. 5.27, strategic decisions about the cancellation and spinning of trains tend to be made in response to major disruptive incidents, to which it may be difficult to apply the types of scheduling approaches already described, and the use of contingency plans may typically be the most appropriate response. However, there may be situations where a trade-off has to be made between operating as many services as is practicably possible, and reducing the delay that inevitably occurs as a result of demand exceeding practical capacity. In such situations, it would be useful to be able to directly compare the disbenefit of cancelling or spinning a service with the cost of the delay incurred by maintaining the service. It would therefore be useful to establish 'proxy delay values' for cancelling or spinning different services. As in the case of train weightings, it may be possible to derive such values from the penalties already imposed on train operators for services which are cancelled or are terminated short of their scheduled destinations.



## 6.0 CONCLUSIONS

- 6.1 It is shown in Chapter 2 that, while the rail mode has characteristics which are highly favourable in some sectors of the transport market, it also has some unique disadvantages in terms of its inherent operational inflexibility and of the propensity for the effects of disruptive incidents to spread widely and rapidly across railway networks. Furthermore, it is argued that rail users are perhaps likely to be less understanding and tolerant of disruptions to rail services than of those to other modes.
- 6.2 For both these reasons, it is important that the likelihood of the occurrence of disruptive incidents be minimised as far as is reasonably and practicably possible, and that such incidents as do occur are dealt with in as effective a manner as possible, so as to reduce their impacts on services and users. The reduction of incident frequency is best achieved by investment in system reliability, while investment in system capacity and redundancy improves the scope for dealing with such incidents as do occur, and reducing the extent and severity of the disruption and delay that occurs as a result.
- 6.3 Given the recent and current state of the UK railway system, the clear priority has been and is for investment in reliability of infrastructure and rolling stock, with system and capacity enhancements in second place. In any case, such enhancements, and particularly the provision of redundancy, may be difficult to justify economically. Until reliability improves, but also generally, it is clearly advantageous to deal with disruptive incidents as effectively as possible, and the need to do this has been noted by the Strategic Rail Authority. This need is emphasised by the fact that the secondary, or reactionary, delay resulting from initial disruptive incidents has increased in recent years. It is claimed that various existing control systems have the capability to deal with service perturbations, but the details of how they perform the task are not in the public domain, and it is therefore impossible to achieve meaningful assessments of, and comparisons between, their respective capabilities.

- 6.4 As implied above and noted in Chapter 3, a useful measure of the effectiveness of efforts to deal with disruptive incidents is the extent of the resulting delay, and the degree to which it is reduced by comparison with the ‘do nothing’ situation. The calculation of delay for re-scheduled trains is straightforward; it may be less so for extensively re-routed services, particularly if they do not serve some of their scheduled stations or stopping points. Similarly, it is difficult to determine comparative values of delay for ‘spun’ or cancelled services; in any case, decisions to take such actions are strategic in nature, and beyond the normal scope of regulation activity. However, it may be possible, and would be useful, to incorporate such decisions into a more general decision-making tool by adopting proxy delay values for such decisions.
- 6.5 In any case, in order to identify delay, and to measure its extent in the cases of trains which are neither ‘spun’ nor cancelled, the scheduled and actual journey times of trains are compared. Frequent such comparisons may be helpful in identifying the initial onset of delays, thus triggering regulatory responses in good time. The theoretical journey times of trains depend on a range of factors relating to rolling stock and infrastructure characteristics, but may be reliably and quickly calculated by computer. The development of a spreadsheet model for performing such calculations comprised one of the major elements of the author’s EngD research activity. The origins, underlying principles and development of the model are described in Chapter 3, and a more detailed technical account is provided in Technical Appendix A. Obtaining data for the validation of the model proved difficult, but it produced very similar results to a textbook example which included consideration of locomotive power and tractive effort, train resistance, and varying grade resistance. The model has already been used by Arup, the Industrial Sponsor, and further potential improvements to it have been identified.
- 6.6 The underlying principles and development of a more general model for the simulation of railway operations are described in Chapter 4. The general characteristics of this model were specified by Arup, with a view to enabling ‘first pass’ assessment of reasonably simple, small-scale analyses of railway operations. In addition to this application, such a model is also useful for the

simulation of responses to disruptive incidents, such as different approaches to the regulation of the movements of trains through congested routes, as described in Chapter 5 of this thesis. In its current state of development, the general model is less ‘complete’ than the spreadsheet-based model described in Chapter 3. However, the fundamental data structures have been established for defining network, signalling and timetable parameters, and a visual display of the modelled network, signals and trains has been developed, with the movements of trains through the network and their interactions with signals being explicitly shown. A means of calculating and displaying the delay experienced by trains has also been developed. The work done to date provides the basis for the further development of the model.

- 6.7 Returning to the original and overriding research theme of disruption management, three main options are identified in Chapter 5 for the regulation of trains in the face of temporary route capacity reductions: re-routeing and/or re-scheduling, ‘spinning’ trains (i.e. turning them back before they have reached their scheduled final destinations), and cancelling them. The main objective of the regulation of disrupted rail services in Britain is the reduction of overall delay, with detailed regulatory decisions being made on the basis of passenger train headcode classes and/or the comparative times required for trains to be routed ahead of one another. This approach has significant limitations in complex, multiple-train situations, and the use of headcodes is not directly applicable to freight trains. A ‘richer’, more detailed measure of the comparative importance of different trains is needed, if an approach to minimising overall train and passenger/freight delay (and resolving the potential conflict between these two objectives) is to be developed and implemented.
- 6.8 The general consequence of such perturbations is a reduction of route (and network) capacity, resulting in increased delay as the route in question and the adjacent elements of the network are operated closer to or beyond their residual capacity. In such circumstances, the effective regulation of trains seeking to use the remaining capacity is crucial. However, such conditions naturally result in ‘fire-fighting’ and are unlikely to facilitate optimal regulation; combined with the speed with and extent to which such disruption can propagate through a

network, this again emphasises the importance of utilising the best possible techniques in such circumstances.

- 6.9 A review of Network Rail's existing approaches to the regulation of disrupted railway services has indicated various shortcomings in the determination of train priorities and in the consideration of multiple (i.e. more than two) trains. The application of existing scheduling techniques has been identified as a potentially promising way forward, and has been demonstrated by means of two fairly simple examples of congested network conditions. Questions remain as to the extent of the network area to which such techniques can be applied, but it is concluded that they may potentially be usefully applied to the network in the area immediately surrounding a capacity-constraining bottleneck, by 'looking ahead' over relatively short periods of time, and updating the assessment of options as trains move through the area under examination so as to take account of the changing relative priorities of trains in the queues on the approaches to the bottleneck.
- 6.10 In order to carry forward the work described in the foregoing material, a strategy and a programme have been agreed within Arup for the ongoing development of the general simulation model. If the opportunity occurs and/or the need arises, improvements may also be made to the spreadsheet-based model for the calculation of train journey time. Further research into disruption management needs and techniques are also being undertaken, at the University of Southampton and in conjunction with Rail Research UK. This includes the further development of scheduling techniques to handle the minimisation of tardiness as well as lateness, and, if possible, to incorporate such issues as the maintenance of connections between trains.

## REFERENCES

- Armstrong, J.H. (1998), *The Railroad, What It Is, What It Does* (4<sup>th</sup> ed.), Omaha: Simmons-Boardman
- Arup Transportation (1993), *Arup Transportation Technical Note: Arup Run Time Model*, Internal Document.
- Bagwell, P. (1974), *The Transport Revolution from 1770*, London: Batsford
- Bagwell, P., Lyth, P. (2002), *Transport In Britain*, London: Hambledon and London
- BBC News (2001), *When road meets rail* [online], BBC. Available from <http://news.bbc.co.uk/1/hi/uk/1507516.stm> [Accessed 14 August 2004]
- BBC News (2002), *Labour strife disrupts Pacific trade* [online], BBC. Available from <http://news.bbc.co.uk/1/hi/business/2289667.stm> [Accessed 13 August 2004]
- BBC News (2003), *Drought-hit farmers plead for aid* [online], BBC. Available from <http://news.bbc.co.uk/1/hi/world/europe/3083625.stm> [Accessed 13 August 2004]
- Bureau of Transport and Regional Economics (2003), *Australian Transport Stats* [online], Bureau of Transport and Regional Economics. Available from <http://www.btre.gov.au/docs/trnstats03/ATS.pdf> [Accessed 13 August 2004]
- Conway, R.W., Maxwell, W.L., Miller, L.W. (1967), *Theory of Scheduling*, Mineola: Dover
- Dalla Chiara, B., Gačanin, E. (2004), *Railway Transport Systems*, Torino: Politecnico di Torino
- De Fontgalland, B. (1984), *The World Railway System*, Cambridge: University Press
- Department for Transport (2003a), *Transport Statistics Great Britain (29<sup>th</sup> ed.): Passenger transport by mode (historical)* [online], Department for Transport. Available from [http://www.dft.gov.uk/stellent/groups/dft\\_control/documents/contentservertemplate/dft\\_index.hcst?n=9045&l=4](http://www.dft.gov.uk/stellent/groups/dft_control/documents/contentservertemplate/dft_index.hcst?n=9045&l=4) [Accessed 13 August 2004]
- Department for Transport (2003b), *Transport Statistics Great Britain (29<sup>th</sup> ed.): Domestic freight transport: by mode (historical)* [online], Department for Transport. Available from

[http://www.dft.gov.uk/stellent/groups/dft\\_control/documents/contentservertemplate/dft\\_index.hcst?n=9045&l=4](http://www.dft.gov.uk/stellent/groups/dft_control/documents/contentservertemplate/dft_index.hcst?n=9045&l=4) [Accessed 13 August 2004]

Department of the Environment, Transport and the Regions (1998), *A New Deal for Transport: Better for Everyone*, The Stationery Office Ltd.

De Vaus, D.A. (1996), *Surveys in Social Research* (4<sup>th</sup> ed.), London: UCL Press

Docklands Light Railway Ltd. (2002), *Technology* [online], Docklands Light Railway Ltd. Available from: <http://www.tfl.gov.uk/dlr/about/technology.shtml> [Accessed 4 September 2004]

Eurostat (2003), *Panorama of transport - Statistical overview of transport in the EU – Data 1970-2001 – Part 2 (PDF)* [online], European Communities. Available from [http://europa.eu.int/comm/eurostat/Public/datashop/print-product/EN?catalogue=Eurostat&product=KS-DA-04-001-\\_-N-EN&mode=download](http://europa.eu.int/comm/eurostat/Public/datashop/print-product/EN?catalogue=Eurostat&product=KS-DA-04-001-_-N-EN&mode=download) [Accessed 13 August 2004]

Faith, N. (1993), *Locomotion*, London: BBC Consumer Publishing

Faith, N. (2000), *Derail: Why Trains Crash*, London: Channel 4 Books

Ford, R., Haydock, D. (1992), 'Signalling and Timetabling', in *Planning Passenger Railways* (Harris, N.G., Godward, E.W., eds.), Glossop: Transport Publishing Company

Fox, P. and Pritchard, R., (2003), British Railways Pocket Book No. 1: *Locomotives* (46<sup>th</sup> ed.), Sheffield: Platform 5 Publishing.

French, S. (1982), *Sequencing and Scheduling: An Introduction to the Mathematics of the Job-Shop*, Chichester: Ellis Horwood

Glover, J. (1999), *Railway Operations*, Hershman: Ian Allan Publishing

Goddard, S. B. (1994), *Getting There*, New York: Basic Books

Goodman, C., Takagi, R. Dynamic re-scheduling of trains after disruption, in *Computers in Railways IX: proceedings of Comprail 2004, Dresden, 17 – 19 May 2004* (Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G. and Sone, S., eds.), WIT Press, 2004, pp765-774

Hall, S. (2001), *Modern Signalling Handbook*, (3<sup>rd</sup> ed.), Hershman: Ian Allan Publishing

Hare, T.B. (1927), *British Railway Operation*, London: Modern Transport Publishing

- Hare, T.B. (c1931), *Practical Railway Operating*, London: Modern Transport Publishing
- Harris, N. (1992), 'Punctuality and Performance', in *Planning Passenger Railways* (Harris, N.G., Godward, E.W., eds.), Glossop: Transport Publishing Company
- Harris, N.G. (2003a), 'Timetabling and Rostering', in *Newcastle University TORG Lecture Course: Railway Management, Economics and Planning* (Harris, N.G.), Newcastle-upon-Tyne: University of Newcastle-upon-Tyne
- Harris, N.G. (2003b), 'Track and Signalling Infrastructure Design', in *Newcastle University TORG Lecture Course: Railway Management, Economics and Planning* (Harris, N.G.), Newcastle-upon-Tyne: University of Newcastle-upon-Tyne
- Hay, W.W., (1982), *Railroad Engineering* (2<sup>nd</sup> ed.), New York: Wiley
- Institution of Railway Operators (2004), *Railway Operating Principles Module 1*, Rail Training International
- Jack, I. (2001), *The Crash That Stopped Britain*, London: Granta Books
- Lamb D.R. (1941), *Modern Railway Operation*, (3<sup>rd</sup> ed.), London: Pitman
- Lloyd's Register Rail (2004), *Definition of abnormal and degraded working* [online], Rail Safety & Standards Board. Available from <http://www.rssb.co.uk/pdf/reports/research/T011%20Definition%20of%20abnormal%20and%20degraded%20working%20-%20Report.pdf> [Accessed 20 January 2005]
- Lu, A. (1996-98), *Ever wondered what the headcode panel means?* [online], Alex Lu. Available from <http://www.lexcie.zetnet.co.uk/headcode.htm> [Accessed 19 August 2004]
- Martin, P. (1999), *Train Performance and Simulation* [online], INFORMS-CS. Available at <http://www.informs-cs.org/wsc99papers/188.PDF> [Accessed 28 April 2004]
- Morlok, E.K., (1978), *Introduction to Transportation Engineering and Planning*, New York: McGraw-Hill
- Nash, C., Coulthard, S., Matthews, B. (2004), 'Rail track charges in Great Britain – the issue of charging for capacity' in *Transport Policy* [online], Elsevier. Available from [http://www.sciencedirect.com/science?\\_ob=MIimg&\\_imagekey=B6VGG-4BWYD4M-1-](http://www.sciencedirect.com/science?_ob=MIimg&_imagekey=B6VGG-4BWYD4M-1-)

[S&\\_cdi=6038&\\_orig=browse&\\_coverDate=03%2F10%2F2004&\\_sk=999999999&view=c&wchp=dGLbVtz-zSkzk&\\_acct=C000010399&\\_version=1&\\_userid=126770&md5=bd5e28c30f537d4487ff1303b43ffbf2&ie=f.pdf](http://www.networkrail.co.uk/cache/2b63e6f4e0474b8cba7c7d7590000bcf.pdf) [Accessed 24 August 2004]

National Economic Research Associates, Symonds Group (2000), *The Standalone Cost Of Freight Access* [online], National Economic Research Associates, Symonds Group. Available from

<http://www.networkrail.co.uk/cache/2b63e6f4e0474b8cba7c7d7590000bcf.pdf>

[Accessed 19 August 2004]

National Safety Council (2000), *Potential Impacts of Climate Change and El Niño in the Mississippi Basin* [online], National Safety Council. Available from

<http://www.nsc.org/ehc/jrn/weather/mississi.htm> [Accessed 13 August 2004]

Network Rail (2003), *2003 Technical Plan. Section 10: Operational Performance* [online], Network Rail. Available from

<http://www.networkrail.co.uk/cache/da746ddd58f844f3b1f7f212c47273e6.pdf>

[Accessed 31 August 2004]

Network Rail (2004a), *2004 Technical Plan, Section 7: Operational Delivery* [online], Network Rail. Available from

[http://www.networkrail.co.uk/Documents/bus\\_plan\\_2004/S07%20-%20Operational%20Delivery.pdf](http://www.networkrail.co.uk/Documents/bus_plan_2004/S07%20-%20Operational%20Delivery.pdf) [Accessed 19 August 2004]

Network Rail (2004b), *2004 Technical Plan. Section 10: Operational Performance* [online], Network Rail. Available from

[http://www.networkrail.co.uk/Documents/bus\\_plan\\_2004/S10%20-%20Operational%20Performance.pdf](http://www.networkrail.co.uk/Documents/bus_plan_2004/S10%20-%20Operational%20Performance.pdf) [Accessed 1 September 2004]

Nock, O.S. (ed.) (1966), *Single Line Railways*, Newton Abbot: David & Charles

Nock, O.S. (ed.) (1980), *Railway Signalling*, London: A & C Black

Office of Rail Regulation (2004a), *Glossary* [online], Office of Rail Regulation.

Available from <http://www.rail-reg.gov.uk/server/show/nav.001002> [Accessed 1 September 2004]



- Office of Rail Regulation (2004b), *The Network Code* [online], Office of Rail Regulation. Available from [http://www.rail-reg.gov.uk/upload/pdf/tac\\_allparts\\_aug04.pdf](http://www.rail-reg.gov.uk/upload/pdf/tac_allparts_aug04.pdf) [Accessed 22 August 2004]
- Pachl, J. (2002), *Railway Operation and Control*, Mountlake Terrace: VTD Rail Publishing
- The Permanent Way Institution (1993), *British Railway Track* (6<sup>th</sup> ed.), Stoke-on-Trent: The Permanent Way Institution.
- Pinedo, M. (2002), *Scheduling* (2<sup>nd</sup> ed.), Upper Saddle River: Prentice Hall
- Simmons, J., Biddle, G. (eds.) (1997), *The Oxford Companion To British Railway History*, Oxford: University Press
- Samuel, H. (1961), *Railway Operating Practice*, London: Odhams Press
- Schivelbusch, W. (1986), *The Railway Journey*, Leamington Spa: Berg
- Schmid, F. (2003), 'The History of Rail Freight and Today's Environment', in *Planning Freight Railways* (Harris, N.G. and Schmid, F., eds.), London: A & N Harris
- SPG Media Limited (2004), *CHANNEL TUNNEL RAIL LINK EXTENSION PROJECT, UNITED KINGDOM* [online], SPG Media Limited. Available from <http://www.railway-technology.com/projects/chunnel/> [Accessed 4 September 2004]
- Strategic Rail Authority (2002), *Capacity Utilisation Policy Consultation* [online], Strategic Rail Authority. Available from [http://www.sra.gov.uk/sra/publications/consult\\_docs/2002\\_09\\_05/cup\\_consultation\\_document.pdf](http://www.sra.gov.uk/sra/publications/consult_docs/2002_09_05/cup_consultation_document.pdf) [Accessed 6 December 2002]
- Strategic Rail Authority (2003), *Network Utilisation Strategy* [online], Strategic Rail Authority. Available from <http://www.sra.gov.uk/pubs2/stratpolplan/cup1/nus> [Accessed 14 August 2004]
- Strategic Rail Authority (2004), *European Signalling System Could Reduce Train Delays by up to 20%* [online], Strategic Rail Authority. Available from [http://www.sra.gov.uk/news/2004/Folder.2004-03-29.7357331632/European\\_Signalling\\_System](http://www.sra.gov.uk/news/2004/Folder.2004-03-29.7357331632/European_Signalling_System) [Accessed 4 September 2004]

Tunley, J. Adhesion management for ERTMS operation, in *WCRR 2003: proceedings of The World Congress on Railway Research, Edinburgh, 28 September – 1 October 2003*, Immediate Proceedings Ltd., 2003, pp1463-1470.

U.S. Census Bureau (2000), *North American Transportation in Figures* [online], U.S. Census Bureau. Available from <http://www.census.gov/econ/www/natf/english.pdf> [Accessed 13 August 2004]

Vranich, J. (1991), *Supertrains*, New York: St. Martin's Press

White, T. (2003a), *Elements of Train Dispatching Vol. 1*, Mountlake Terrace: VTD Rail Publishing

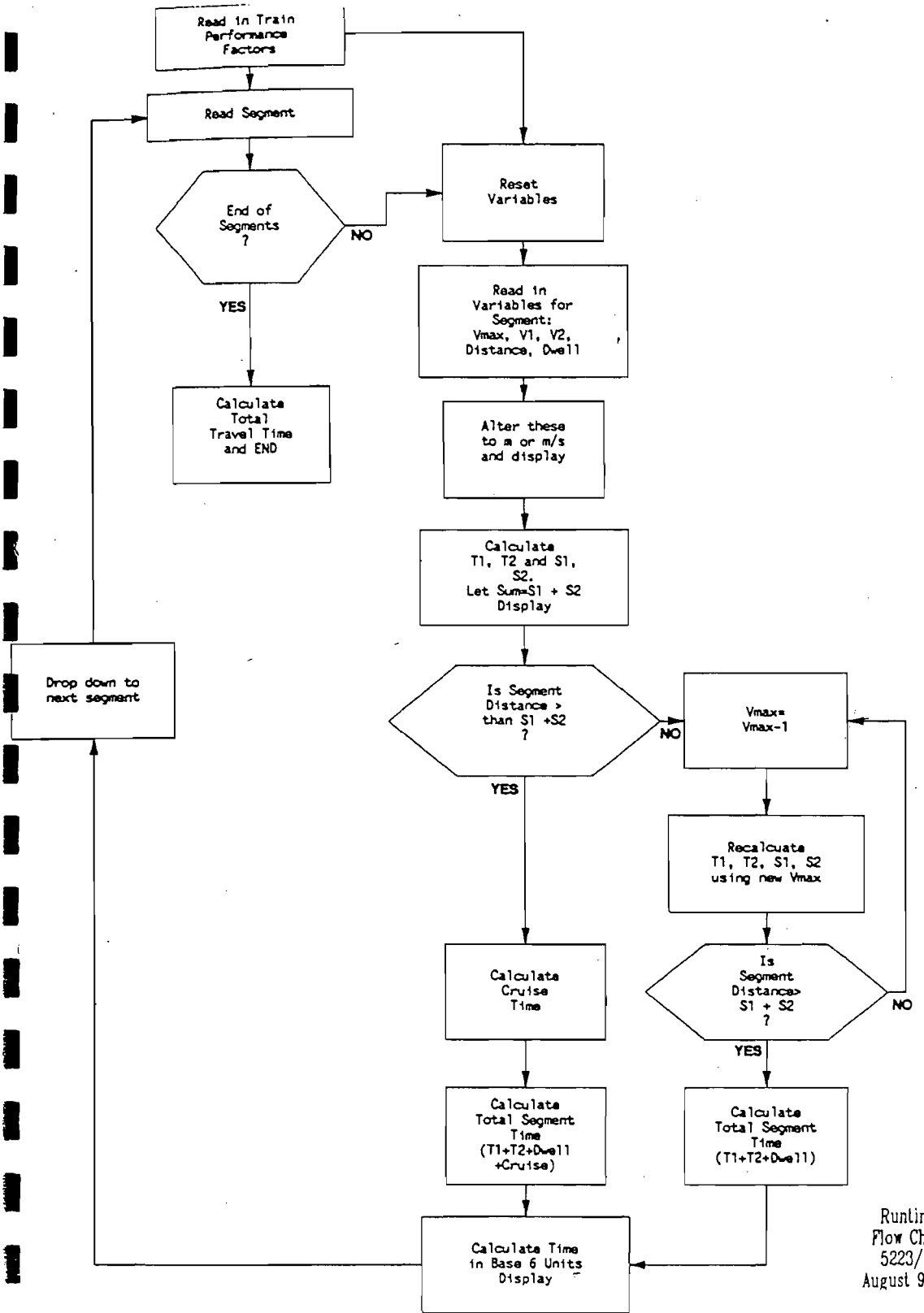
White, T. (2003b), *Elements of Train Dispatching Vol. 2*, Mountlake Terrace: VTD Rail Publishing

Wolmar, C. (2001), *Broken Rails*, London: Aurum Press

Woof, T. (1999, 2001), *Kilo Newtons, kilo Watts, kilometres per Hour* [online], T. Woof. Available from <http://www.twoof.freemove.co.uk/motion1.htm> [Accessed 9 July 2004]

**APPENDICES**

## **Appendix A: Extract from Original Arup RUNTIME Model Documentation**



Runtime  
Flow Chart  
5223/19  
August 93/bs

Figure 2

ARUP RUNTIME : SUMMARY OUTPUT SHEET

Job Name HEARDALE Description: Base Run  
 Job Number 46145 No Line Speed Increases  
 Date 6/ 9/93

Summary Results Total Time: 26 Mins 38 Secs  
 Distance: 19.2 km

Train Factors Type of Train 156 SPRINTER  
 Av Acceleration .5 m/s2  
 Av Deceleration -.5 m/s2  
 Max. Speed 97. kph

Summary Breakdown

Description	Seg No.	Input Data				Enough Distance ? (1)	RESULTS		Cum. Time		New Vmax kph
		Dist km	V1 kph	V2 kph	Vm DWELL kph sec		Clock Time Min Sec	Min Sec			
DARLINGTON	1	.52	0	32	40. 0	YES	0 58	0 58			
	2	.56	32	32	48. 0	YES	0 44	1 43			
	3	.64	32	24	32. 0	YES	1 12	2 55			
NORTH ROAD	4	.34	24	0	32. 60	YES	1 47	4 43			
	5	.52	0	32	32. 0	YES	1 7	5 50			
	6	.87	32	32	32. 0	YES	1 37	7 28			
	7	2.2	32	48	72. 0	YES	1 58	9 27			
HEIGHTINGTON	(8)	1.1	48	72	48. 0	YES	1 8	10 35			
	9	.34	72	72	72. 0	YES	0 17	10 52			
	10	2.2	72	0	72. 30	YES	2 41	13 34			
	11	.24	0	54	54. 0	YES	0 31	14 5			
AYCLIFFE	12	1.8	54	0	72. 60	YES	2 51	16 57			
	13	1	0	72	72. 0	YES	1 9	18 7			
	(14)	2.0	72	48	72. 0	YES	1 40	19 47			
SHILDON	15	.20	48	0	48. 60	YES	1 28	21 16		35.75997	
	16	.60	0	48	48. 0	YES	0 58	22 15			
	17	.16	48	48	48. 0	YES	0 12	22 27			
	(18)	1.1	48	48	48. 0	YES	1 24	23 51			
BIS AUCKLAND	19	2.7	48	32	72. 0	YES	2 24	26 15			
	20	.12	32	0	32. 0	YES	0 22	26 38		27.67999	

NOTES

- \*V1 = Start Speed
- \*V2 = End Speed
- \*Vm = Max Speed
- \*(1) = Sufficient Distance to Reach Vmax

Table 1

ARUP RUNTIME : SUMMARY OUTPUT SHEET

Job Name WEARDALE Description: Final Test Case  
 Job Number 46145 Line Speed Increases Test 1  
 Date 6/ 9/93

Summary Results Total Time: 21 Mins 34 Secs  
 Distance: 19.2 km

Train Factors Type of Train 156 SPRINTER  
 Av Acceleration .5 m/s2  
 Av Deceleration -.5 m/s2  
 Max. Speed 97. kph

Summary Breakdown

Description	Seg No.	Input Data				Enough Distance ? (1)	RESULTS		Cum. Time		New Vmax kph
		Dist km	V1 kph	V2 kph	Vm DWELL kph sec		Clock Time Min Sec	Min Sec			
DARLINGTON	1	.52	0	40	40. 0	YES	0 57	0 57			
	2	.56	40	48	48. 0	YES	0 42	1 40			
	3	.64	48	56.3	56. 0	YES	0 41	2 21			
NORTH ROAD	4	.34	56.3	0	56. 60	YES	1 37	3 58			
	5	.52	0	72.4	72. 0	YES	0 45	4 44			
	6	.87	72.4	72.4	72. 0	YES	0 43	5 27			
	7	2.2	72.4	72.4	72. 0	YES	1 49	7 17			
	8	1.1	72.4	96.5	97. 0	YES	0 (40)	7 58			
HEIGHTINGON	9	.34	96.5	96.5	97. 0	YES	0 12	8 11			
	10	2.2	96.5	0	97. 30	YES	2 19	10 31			
AYCLIFFE	11	.24	0	54	54. 0	YES	0 31	11 2			
	12	1.8	54	0	72. 60	YES	2 51	13 53			
	13	1	0	96.5	97. 0	YES	1 4	14 57			
SHILDON	14	2.0	96.5	48	72. 0	YES	1 (35)	16 31			
	15	.20	48	0	48. 30	YES	0 58	17 29	35.75997		
	16	.60	0	56.3	56. 0	YES	0 54	18 24			
	17	.16	56.3	56.3	56. 0	YES	0 10	18 34			
BIS AUCKLAND	18	1.1	56.3	96.5	97. 0	YES	0 46	19 21			
	19	2.7	96.5	38	97. 0	YES	1 51	21 12			
	20	.12	38	0	38. 0	YES	0 21	21 34	27.67999		

NOTES

- \*V1 = Start Speed
- \*V2 = End Speed
- \*Vm = Max Speed
- \*(1) = Sufficient Distance to Reach Vmax

Table 2

**Appendix B: Results of Train Performance Calculations by W.W. Hay**



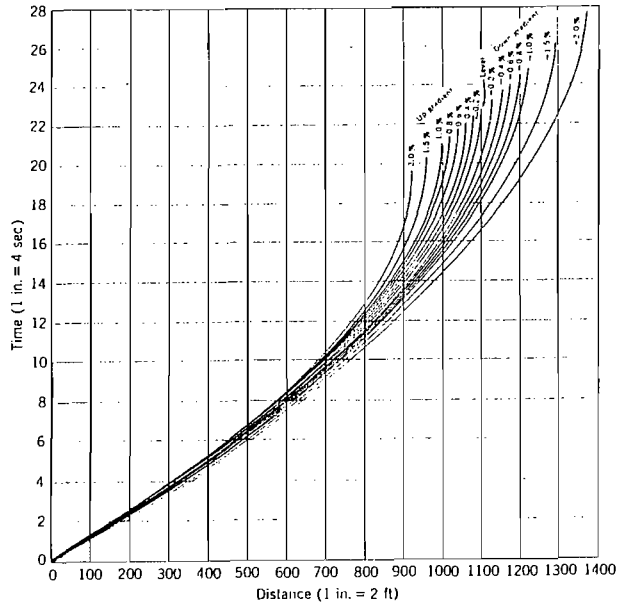


Figure 10.10. Time-distance curves for braking on various grades (computed for four 2000-hp diesel-electric units pulling 50 70-ton cars).

**PROBLEM EXAMPLE**

The use of acceleration, deceleration, and braking distance curves can best be understood from a problem example.

**10. Train Performance Problem**

A fragment of a train performance study using the procedures of this chapter appears in the profile and curves of Figure 10.11. Superimposed on the physical profile of a line are the corresponding portions of speed-distance and time-distance curves for the performance of four 2000-hp diesel-electric units pulling 50 70-ton freight cars. The basic curves have already been computed and presented in Figures 10.3 and 10.4. On 12,000 ft of level grade the train accelerates to a speed of 47.4 mph. The corresponding portion of the speed-distance curve for

a 0.0% grade is transferred to the profile by tracing or replotting. There is thus an advantage in using the same scales on the speed-distance curves as are used on the physical profile.

On attaining a speed of 47.4 mph, the train encounters an ascending 0.6% grade that acts to decelerate the train along the corresponding speed-distance curve from 47.4 mph to some lower speed. After decelerating to 42 mph, the 0.6% grade gives way to one of 0.2%, and the train again accelerates. At a speed of 47 mph the brakes are applied and the braking distance curve (Figure 10.10 for a 0.20% grade) is followed to a stop. The combination of these several portions of speed-distance curves gives a graphic presentation of the demand on the locomotive and its capabilities at every point in the run. The profile shows where limiting conditions occur and where the locomotive is not working to its maximum capacity. Note that where the locomotive is maintaining a constant speed, the speed-distance profile would become a horizontal straight line.

A speed-time curve is plotted in the same way, using the data from Figure 10.4. This curve plotted on the physical profile shows the overall time to traverse the route as the maximum ordinate distance. The time between any two points may also be read graphically. Note that if the train of Figure 10.11 had stopped en route, the continuing time during the stop would have increased on a straight vertical line.

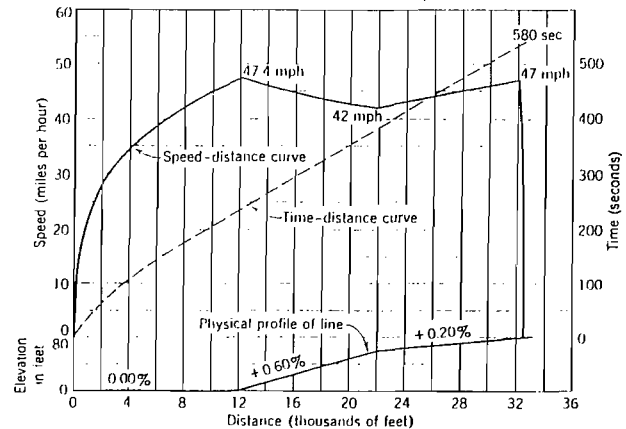


Figure 10.11. Speed-distance and time-distance curves showing the performance of four 2000-hp diesel-electric units pulling 50 70-ton cars).