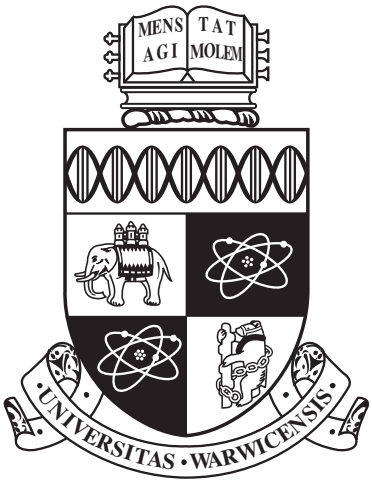

Data-driven Network Analysis and Delay Study of the British Railway

by

Andrew Hilditch

Thesis

Submitted to the University of Warwick

for the degree of

DOCTOR OF PHILOSOPHY

Mathematics of Real World Systems

November 2022

THE UNIVERSITY OF WARWICK

Contents

Acknowledgments

Completing this thesis took a long time and encountered many issues along the way. Fortunately, I was not alone in my work. I would like to thank my supervisor Professor Colm Connaughton because if I had a different supervisor the work would not have gotten to this stage and would likely have had no content at all.

I would also like to thank my industrial supervisors from Thales, a French engineering firm. My first supervisor was Tim Cornah and he taught me the value of a data first approach and supported me throughout. Dr Dave Fidal also lent me his support with his expert rail knowledge from many years working with the railway data. The last of my industrial supervisors was Ian Bradshaw who I would like to thank also for his knowledge and making time for me to meet with him regularly. Special thanks also to Darren Barnard for some of the data and for answering my questions many of which were poorly worded and to Matt Alway who gave some of the schedule information. Joanna Thornton also helped me with staying on course for meeting industry obligations.

Thanks should also go to the Mathematics of Real-World Systems CDT at Warwick university and especially Professor Michael Tildesley and Jade Perkins for meeting with me in the final months to give me guidance as to the best way forward for me to take in completion.

I would like to thank the EPSRC and Thales for funding this research with which this work would never have been done. Thanks as well should go to my supporting family who helped me as I worked from home throughout the Covid-19 pandemic.

Thanks also to the members of the transport group in the CDT which I

attended where I saw many interesting talks and tried to give many of my own.

I am grateful also to the various communities of railway enthusiasts on the internet and especially the open rail wiki, open train times and the railway codes website maintainers. These were a good set of websites to look at and learn from.

Declarations

This work has been written by me and has not been submitted for any other degree, award or qualification. There was no collaboration beyond that of the supervision from Professor Colm Connaughton. None of this work has been previously published.

Abstract

In this thesis we look at the British railway network and perform data-driven network analysis on publicly available data. We perform centrality analysis as well as basic network theory methods on different levels of the network. The analysis allows us to see that the historic hubs can be found from current data, and we also located some of the modern hubs. We then move onto an in-depth case study of a local line where we examine bottlenecks and learn that the section of line nearest to London has the most issues and discover that turn-around times at terminuses pose an obstacle to catching back up from delay. We later propose a formula to estimate outbound delay with respect to inbound delay. Two Random Forest Regression methods are used in this thesis. The first is used to guess how delayed a train will be at a station and is used to find the feature importance measures. We find that the time of day and the station being predicted are the two factors with the greatest contribution. The second regression method is used to predict travel times to stations. This method is compared to a baseline prediction and outperforms it easily. Work is also done to examine train interactions and knock-on delay. We propose that trains on the line can form "concertinas", a relationship where trains move closer and then further apart, this delays the rear train. These relationships are studied in the form of interval graphs.

Chapter 1

Introduction

In this thesis we will be examining various aspects of railway analysis. Chapter two primarily examines network theory as applied to different levels of the network (national, regional, local). This work is useful since the current owners of Great Britain railway have difficult to maintain information as to how the network is connected at a fine grain level of detail. Our work can also be used to keep up with changes made during re-signalling work. Chapter three looks at bottlenecks and other infrastructure issues that disrupt planned travel at a local level. This has obvious applications to real world use as reliability is of increasing concern on the Great Britain railway [62]. Chapter four's work is on train interactions and proposes methods to predict travel times on the railway also using a more localised view. This work could better inform train operators and signalmen as to how the existing trains will fare later in their journey. The chapter also addresses turn-around times at terminals and how delay propagates through the network.

The U.K rail network is the oldest in the world, the first section opened to the public in 1825 with the Stockton and Darlington railway [11]. The network has a long history of development with the impact of construction from the 19th century still being felt today [27]. Initially, the network was made up of many different private lines that eventually were connected to become a national railway network [66]. This history led to railway hubs being in the "wrong" locations from a population perspective such as Crewe, Swindon, Eastleigh and Ashford [11] instead of Manchester, Birmingham and Leeds for example. They form part of the current network of hubs across the GB railway along with more recent ones like Birmingham New Street.

After being nationalised in 1947, the railway was subsequently re-privatised with the last public service being run in 1997 [95]. This continued until recently

when the East Coast Mainline and the Northern Rail franchises were transferred back to public ownership [6] in 2018 and 2020 respectively. That was until the Covid-19 pandemic where the government had to intervene due to passenger numbers dropping to 5 percent of pre-covid amounts [97]. Liddell [61] discusses the reasons behind the loss in passenger numbers and attributes this decrease to U.K. government messaging, post-lockdown restrictions, an adaptation to working from home and customer sentiment. Coppola et al [15] describe the loss of passenger capacity from social distancing in greater detail. Our work in this thesis focuses on data from 2019 therefore the Covid-19 pandemic will not be influencing the work shown from now on. Keywords and phrases that are later referred to in this thesis will have longer explanations when they appear in the subsequent chapters. There is a summary of abbreviations at the end of this work.

Use of network theory [68] in the analysis of railways is well documented. One application is the study of resilience against delay. Pagani et al [71] uses trophic coherence to model the resilience of the network during morning rush hours into Greater London. Trophic coherence is a concept originating in food webs that quantifies the hierarchical structure of a directed network by assigning levels to nodes. Feng et al [29] describe the use of temporal networks to look at resilience of the Chinese railway network as well as that of Paris. Temporal networks are a type of network where edges can be active only at certain points in time. Network theory can also be used to study accidents that occur on the railway. An example of this can be seen from Jintao et al in their 2019 paper [64] where the authors use accident reports from various national bodies around the world to construct a railway operational accident causation network. This network is then used to find which hazards cause the most accidents. Another paper that links accidents to Network theory is Lam and Tai in 2020 [55], where they study accidents in Japan by plotting a network of incidents. These networks show the links between causes of incidents and the outcomes from the situation. Other papers discuss more general methods of application, for example [21] where the authors use complex network theory to map the growth of the Kuala Lumpur rail transit network, [92] where the authors are examining different metrics of the Russian railway network with centrality measures and [53], this last paper looks at directed network of freight in Eastern Europe and solves it in a game theory format. We apply network theory in chapter two of this thesis to the GB railway at local and national levels.

Bottleneck analysis and utilisation of the network are closely related concepts. A bottleneck forms when a section of the network is over-utilised. This is an area of research that will be addressed in chapter three of this thesis. First, we will

address some literature on utilisation rates, Laroche in 2013 [58] proposes a method to estimate these rates. It uses real data from the French railway system to model utilisation rates as a function of the capacity of the line. Armstrong and Preston in 2017 [3] also examine utilisation and bottlenecks using data from a previous project to categorise locations. They then look at train arrivals in differing periods of time to estimate the utilisation rate for each location.

Chen et al in 2015 [12] looks at rescheduling trains as they approach bottlenecks to increase the throughput of trains. This is done using a mixed-integer programming model to solve the problem according to a set of constraints. D'Ariano et al in 2008 [28] also look at flexible timetables to manage bottlenecks and they rely on more active management during operations with less advanced planning. They do this by adding more gaps into the schedule, unfortunately this reduces throughput and overall passenger capacity.

In chapter three of this thesis, we will utilise random forest regression. In 2004 Segal [80] evaluates random forest regression and discusses the method in detail. He says that "the method entails using an ensemble of trees, where each tree in the ensemble is grown in accordance with the realization of a random vector". The method generates a number of random trees that each make predictions based on training data. These predictions are then averaged to give an overall prediction. His approach prioritises reducing overfitting, which is when a statistical model conforms to a data set too much and then becomes incapable of predicting future values. Segal examines a repository of data and suggests improvements to the regression methods used there. Along explanation of random forest regression is found in chapter three. Yaghini et al in 2013 [101] explores using a machine learning method to predict train delay. It uses a multilayer perceptron neural network which is then fitted to training data before being tested and validated. The conclusion reached is that the model performed accurately. Grömping compares linear regression to the random forest method in his 2009 paper [38]. He explores the concept of variable performance later in the paper which is of interest to the work later in this thesis. Montgomery et al in 2021 [67] has an extensive look at linear regression and contains examples where the method is implemented. The previous paper is useful since in chapter four of this thesis linear regression is used.

Another important aspect of studying railways is that unlike many other forms of transport the trains operate with a timetable. The analysis of schedules is an area of research that many have examined. Goverde describes timetables in detail in his 2005 thesis [36] before going on to describe how to optimise these timetables. His later work published in 2007 [37] continues to discuss timetables

in an effort to further increase use of existing infrastructure and improve reliability through the development of a stability analysis tool. Ochiai et al in 2019 [69] analyse drivers' behaviour to prevent drift from the timetable. They do this by using decision trees of driver actions. Artan et al [4] use Markov chains to track transitions between different delay states to attempt to return to the timetable. This helps evaluate timetable robustness. Work on improving timetables is also seen in Khoshniyat et al in 2017 [50] where they discuss the minimum headways in the schedule between services. They scale these headways with the length of the services, again to attempt to improve the robustness of the timetable. Lee et al in 2016 [60] suggested a method to adjust the timetable by using decision trees to find the root cause of delays and then altering later timetables to avoid these issues. Shakibayifar et al in 2017 [84] discuss scheduling the Iranian railway through the use of mixed-integer programming. In chapter four of this thesis, we examine the feasibility of the timetable on the line that we are looking at.

Congestion on the railways is an issue that will always be present and increasingly so as various countries try to get the most capacity out of existing infrastructure. This topic is covered by a substantial amount of research due to the obvious financial benefits of improving the congestion problem. Railway traffic can be split into two general categories, these are freight and passenger. An example of a paper focusing on freight is Gorman in 2009 [35] where the author examines congestion in the US freight system. The main motivation is to reduce delay without building more railroads. This is done by modelling types of delay and their effect on congestion rates. Regarding the passenger focused papers, many of them from developed countries look at Japan. Examples of this are Hibino et al in 2005 [40], Ono et al in 1999 [70] and Yamauchi et al in 2017 [102]. Hibino and Ono's papers are exploring increasing throughput of trains to increase capacity, whereas Yamauchi et al is considering congestion of passengers on the trains. They are doing so by optimising stopping patterns so the paper is of value here. Lindfeldt in his 2015 thesis [63] examines the Swedish railway, also working on getting more capacity from what is an already saturated rail network.

Some other congestion related research papers such as Fullerton et al in 2014 [30] and Gibson et al in 2002 [33] look at different problems related to congestion. Fullerton's paper looks at the additional fuel cost of congestion and Gibson looks at the compensation costs paid by the infrastructure owner due to congestion. This thesis will examine congestion and knock-on delay in chapter four.

An area of interest to my work is train interactions. These can be in the form of knock-on delay or simply being close enough to be considered interacting with one

another. Yuan and Hansen in 2008 [103] describe minimising knock-on delay through optimising buffer times at bottlenecks. This is done by measuring the headway between trains at junction approaches and then using an optimisation model to minimise knock-on delay. Another relevant paper on headway is Landex and Kaas in 2005 [57] where a balance is described between capacity and reliability. The paper is about optimising this balance with respect to the existing fixed-block signals in place. This is interesting since many of the other papers on this topic decline to mention the fixed-block nature of the railways they are studying. Fixed-block signals are a widely used method of signalling that has signals in fixed locations at the trackside that govern train entry to sections of track. They can also inform drivers of trains ahead with instruction to slow down or stop. This is a major issue if not addressed when attempting to apply research to reality. The conclusion suggested is to reduce travel speed by six percent to reduce headway by eleven percent. It is not mentioned that this will increase travel times and therefore negatively affect the passenger experience. The paper instead focuses on the increase to capacity.

Interval graphs are an area that will be explored in chapter four as well. Golumbic in 1985 [34] discusses this type of graph and mentions how many real-world applications feature this approach to graph theory. The author goes on to classify various different types of interval graphs and discusses research in the area. The author proposes the following definition of an interval graph "An interval graph is the intersection graph of a family of intervals on the real line, i.e., to every vertex of the graph there corresponds an interval and two vertices are connected by an edge of the graph if and only if their corresponding intervals have nonempty intersection." Gattass and Nemhauser in 1981 [31] apply the study of interval graphs to pavement analysis, here meaning road surface, in New York city. This is done with interval graphs and different propositions are used and then proved. In their 2003 paper Habib et al [39] discuss common connected components (CCP) of interval graphs. They propose a solution for solving the CCP problem with respect to interval graphs.

The transport logistics industry is a vital part of the modern world. The transport of goods around the world are growing increasingly by the year [24]. The global logistics market was valued at 981 billion euros in 2011 [24]. Many papers have looked at the management of logistics. Ducruet et al in 2012 [26] looked at container freight networks. They examined the growth in container shipping through the use of graph theory. They also compared two different networks of freight ports; this was useful as a case study to compare between graphs. Kim et al [51] stated that shipping liners now demand high performance levels from container ports. This means that ports have to improve throughput. There are many different complexities that must

be handled at a container port such as unreliable arrivals tidal issues and crane usage [51]. This is comparable to the various issues handled at railway stations. Examples of which are late arrivals, platform usage and train cancellations. Many of these approaches use operational research methods [88]. The area of operational research looked at most commonly is optimisation. This looks at getting the most out of the system being examined. In the container port this would be containers unloaded. In the railway station this would be trains and passengers handled. These would be the objective function. The rest of the problem is formulated as constraints. Examples would be only one train in each platform, trains need a minimum of one minute to unload and trains need to stay three berths apart. This problem is then solved to maximise the objective subject to the constraints. Issues with this approach are that too many constraints can make the problem increasingly time-consuming to solve [16]. Multiple objectives can further complicate solutions being found. This approach was not chosen to be used in this thesis because station management was not looked at.

Road traffic is another area that has applicable approaches to railway problems. The main difference between road analysis and the railway is that road traffic is considered to be continuous in most circumstances. Individual cars are hard to track on motorways. Railways have reliable discrete data on trains. Automatic number plate recognition cameras can track individual vehicle journeys but are mostly used for calculating travel times between successive camera locations [78]. Robinson et al [78] suggests that automatic number plate recognition could be used to develop origin-destination pairs. There are ideas that can be taken from road travel that are applicable to railways. Network theory is one such area. Tian et al [91] looked at the urban road network and looked at the capacity of road sections. This approach also looks at utilisation, here called efficiency. This focus on utilisation of existing infrastructure is looked at in chapter three of this thesis. Work on managing excessive traffic is another focus of road traffic analysis. Knorr et al [52] looks at predicting traffic jams before they occur to try to prevent them from occurring. They do this by using smart vehicles to send data and then simulate traffic flow breakdown. Data from individual trains is not made public on Network Rails' datafeeds but this approach is similar to the concertina model in chapter four. Both systems look at precursors to traffic delays.

Air travel is similar to rail travel in its discrete nature and the airports have runways that are similar to the platforms at railway stations. Methods from rail travel could also have useful applications in the study of railways. Jaquillat and Odoni in 2015 [46] looked at airport congestion and optimised usage of runways to

re-arrange flights. This approach, when implemented with a queuing model reduces delay. Wang et al in 2019 [98] looked at how airlines try to reduce delay for their services. This is an interesting approach because most academics study the airport side of the problem. This is done by comparing different styles of network. The case study that we examine in this thesis has only one passenger operator so improvement in delays at any station reduces delay for the operator.

Predicting travel times and arrivals on railways is an extremely useful application of research for train operators. It can allow them to plan better for disturbances that have yet to occur, possibly allowing for better handling of incidents where delay is caused. Jiang et al in 2019 [47] look at this very problem using various prediction methods to predict delay recovery from primary delays. They describe their data in detail and then make an assumption that trains experience only a single delay. This differs from work shown later in this thesis as delay is considered not by delay event but rather as a variable that is constantly in flux. In the aforementioned paper the authors describe sections of the track that have high recovery potential. Importantly, these trains are considered in isolation from one another. They conclude by stating that their model can forecast delay recovery better than other prediction models.

Pongnumkul et al in 2014 [73] look at improving predictions of arrival times for passenger trains using historic travel times. This is similar to the work done in this thesis in chapter four. Their algorithm improves prediction error when compared with a baseline prediction, in this case the scheduled arrival plus the current delay. A baseline algorithm is a comparable existing metric that can be used to show improvements made with the new method. They study the Thai railway and have detailed location information from GPS. Interestingly, the authors seem not to know of National Rail's DARWIN system used in Great Britain for predicting arrival times. They use station travel time predictions which are not as fine-grained as the ones used in this thesis. They also use a k-nearest neighbours model which looks ahead at trains that have recently arrived and uses historical information to inform how much delay the current train will experience subject to a stated formula. Our work later in this thesis takes this idea further by using a regression model to further refine predictions.

In 2020 Huang et al [45] published a paper that looks at improving travel time predictions during disruptions. This is similar in topic to the aforementioned Jiang et al paper in 2019 [47]. Huang's paper aims to improve re-scheduling of train services to reduce the effects of train conflicts. The authors highlight the issue of real-time predictions using machine learning methods. The paper uses a

machine learning model combined with a Kalman filter to attempt to solve the live prediction time issues as well as the historical prediction problems of lacking up-to-date context. Their method improves the live prediction capability of the machine learning method.

This thesis looks at the application of network theory as applied to the Great Britain railway network. The particular focus is on centrality measures to find the hubs. This is being performed with berth-level analysis. Berths are sections of rail track that are used as part of a safety system, this prevents trains from occupying the same section of track. They can vary in length from station berths that are less than a hundred metres in length to rural lines where they are a few kilometres long. However, the length of berths is not well documented [106]. This area is understudied due to it being regarded as just a safety system. Because of this our industrial partner Thales GTS suggested using it to provide more frequent updates of train location and a higher granularity than just examining stations and junctions. This approach, due to its innovative nature, required more effort to process and clean the data. This process is described in chapter three and the code is available online to allow for further research in this area.

During the process of this PhD the Covid-19 pandemic occurred. This has reduced the need to resolve capacity issues on the railway network in Great Britain due to it being a primarily passenger-based railway. Our industrial partner valued the work looking at delay diagnosis seen at the end of chapter three. They said that it validated and added numerical values to justify decisions that had been made in the past. The turnaround time analysis found in chapter four was also of interest and they could see immediate applications to scheduling of trains.

Despite the setbacks for railways in Great Britain this work could still be applied to other railways in the world. It would be most useful for passenger railways though. This is due to the freight industry having less interest in punctuality than the more customer orientated commuter services. The concertina model proposed in chapter four is applicable to all railways that use fixed-block signalling. A similar relationship can be seen on moving-block railways but does not show the same behaviour to the fluctuating distances seen in this thesis.

Chapter 2

Construction of the Digital Model for the Railway Network from Berth Data

2.1 Introduction: Background on Network Rail data

The Great Britain railway network is made up of over 20,000 miles of tracks that span the island of Great Britain. There are over 2,500 stations on this network [76] with only a few run by the track infrastructure main operator, Network Rail. The rest of the stations are managed by various train operating companies (TOCs). These TOCs are chosen through a franchise process and have licenses to run trains on sections of track. However, they must also agree to a set of restrictions such as having to run a certain number of trains at peak and off-peak times.

The network itself is ever changing, seemingly always out of date, even during construction in some cases [20]. This constant changing of infrastructure can confuse operators and slow down maintenance. This failure to track assets can lead to costly additional audits [20]. As a result of this, the ability to manufacture an updated version of track layout and connected topology from existing data could prove invaluable. This could be avoided if all related parties updated a central database of information describing the network. But such a database does not exist for the Great Britain railway network. Dentten [20] states that a digital version of the project, in this case Crossrail, has the same importance as the physical construction itself. In this chapter we will reverse-engineer a digital model of the berth network from publicly available data. Previous work has been done in this area in papers such as Shabelnikov et al [83], where the authors wrote about digital twinning in the

context of railway infrastructure and found that "implementation of digital twins improves economic indicators of work of an object" as well as improved safety and warns of fault situations. Digital twinning is also used by others like Teshima et al in 2014 for improvement of traffic management systems which require a network model to work effectively [89] it is used in this paper for automatic route setting.

The task of constructing the network from data is not as simple as it may seem. There are many granularities of scale that can be chosen. You could for example look at station level connections and consider only that behaviour. Junctions could also be of interest and together they are referred to as TIPLOCs (Timing Point Locations) in the industry. This grouping historically applies to passenger trains, freight uses a different system called STANOXs (Station numbers). Unhelpfully, the relationship between the two is a many-to-many mapping, with most of the differences being in freight depots. The data that we will be using comes in a berth format. Berths are relatively short sections of track that are between less than a hundred metres and a few kilometres long. They are part of the signalling system, mainly used for maintaining a safe distance between trains. Fortunately, for our work, they can be used to track train movements that can allow for the construction of the network.

Not all areas of the GB railway network are digitised or even have electricity. These areas are governed by token passing and semaphores[14]. These systems are mainly used on rural single-line segments. They will not be studied in this work because berth data for these areas is not available. Network Rail continues to operate around 110 token machines, and some more are used on heritage railways[48]. Due to the small scale of token passing and semaphores they have a small impact on railway traffic. The main loss of information is in regards to the connectivity and topology that we lost leaving these areas out. More work could be used to integrate these areas in the future.

Berths have signals at the start and end of them managing the entry of trains into the section of track. These signals prevent trains from being near each other which prevents collisions from occurring. The signals work similar to a traffic light system displaying various information to the driver [100].

2.2 Description of the berth data:

The data itself can be sourced publicly from Network Rail's data feeds[75]. We received assistance from our partner company to get a sample of the data as well as the expertise to supply us with documents that are not available to the general

public. These documents allow for greater understanding of the data and how it is formatted. There are two types of data obtained from these feeds. The first is the berth data which will be the principal focus of this chapter. The second data output is the schedule data which returns the timetable of the train services on the track. This data set is used in part later on in the chapter.

This berth data typically comes in text format consisting of a sequence of messages in string format that are sent when movements occur in real time, below is a sample of the raw data:

```
<CA_MSG>ESCA138513871G29115958</CA_MSG>
<SF_MSG>CCSFBB00115958</SF_MSG>
<SF_MSG>G2SF91F7115959</SF_MSG>
<SF_MSG>SSSF3108115959</SF_MSG>
<SF_MSG>LCSF157F115959</SF_MSG>
<SF_MSG>CASF8A04115959</SF_MSG>
<CA_MSG>BNCA042300359S60115959</CA_MSG>
<CA_MSG>BECAE114E1121W12115959</CA_MSG>
<SF_MSG>T2SF5979115959</SF_MSG>
<CA_MSG>Q1CA041604081Y37115959</CA_MSG>
<SF_MSG>LBSF02B8115959</SF_MSG>
<SF_MSG>WISF6784115959</SF_MSG>
<SF_MSG>G1SF6640115959</SF_MSG>
```

In general, we are only interested in the CA (Berth step) messages since the SF (Signalling update) entries refer to changes of signals [106]. Although the CB (Berth cancel) and CC (Berth interpose) messages have some uses for managing trains in and out of the network, they do not generally represent movements between berths. The CA messages are berth movement records that maintain up to date information about where a train currently is. An example of a CA message can be seen below:

```
<CA_MSG>ESCA138513871G29115958</CA_MSG>
```

The relevant part of the message is between the CA message designations. The content of this record is broken down in the following table:

AreaID	Message Type	From Berth	To Berth	Train Headcode	UTC Time
ES	CA	1385	1387	1G29	115958

Table 2.1: Breakdown of CA message from the Network Rail datafeeds.

The different fields explain the contents of these messages. The area ID says which Train Describer the movement occurred in. Train describers are sections of the network governed by signalmen. The train headcode describes the type of train, the regional destination and the number of services which match both the train type and regional destination. This number starts at "00" and increases by one each time another service with the same first two characters occurs. Another example of this breakdown can be seen in a paper written by Zhao et al in 2016 [105]. Unfortunately, the headcode is not unique. This means that there can be more than one instance of this headcode moving around the network at the same time. It is possible that two trains with the same headcode can be in the same area at the same time. If that occurs, one of the headcodes will instead report with the headcode entry ****. The other fields in the table are self-explanatory.

Berth IDs are a non-unique description of a section of track known as a berth. They are derived from a safety system designed to prevent trains from being too close to one another. The majority of berth IDs consist of four numbers but there are exceptions to this rule. This is best shown with non-standard berth IDs such as HG11 or N112. These are only decipherable with local knowledge akin to that of the signalman from that area. HG is from the Edgeley Train Describer and stands for Hazel Grove which is a nearby area of Stockport. Similarly, the N berths from the CV Train Describer represent Nuneaton, an area north of Coventry [99].

The schedule data contains both the headcode and a more unique form of identification known as the trainUID which is short for train unique identifier. Despite the encouraging name the trainUID is not actually completely unique and there are multiple train classifiers that are more unique. The trainUID refers to a unique service that has a schedule and the trainUID is assigned to only one train per day on the network.

Greater degrees of uniqueness can be had if the same train classifier would not be used on different days or even for different physical trains. One example of this identifier lacking uniqueness is that our data set for a day starts at 4am in the morning and runs until 4am the next day. If a service runs that starts before four and finishes after it, there will be two different services with the same trainUID running on the same day.

Unfortunately, these trainUIDs are not in the berth movements dataset. We have provided a more detailed breakdown of the schedule data in appendix A.

2.3 Creating a network of connected Train Describers

The Great Britain railway network is split up into different Train Describers and each one is managed by a signalman. Each of these Train Describers contains a number of berths ranging from as few as twenty berths to as many as several hundred distinct berths.

The objective of this section is to create a network of adjacent Train Describers thereby allowing checks on whether berth movements between areas are plausible. This created network is also interesting in its own right as a higher-level network for analysis. It shows us which areas of the country experience the most traffic and matches these to conventionally known hubs within the network. Nodes in this network are Train Describers and edges are physical connections between Train Describers.

By tracking headcodes of trains and sorting them chronologically we can track trains moving between describers with adjacent berth movements. We then count how many times in a day a service passes from one train describer to another. To register as an edge in our network at least ten trains had to pass between the Train Describers.

We have made several attempts at displaying this network, some more successful than others. Figure 2.1 shows the connected Train Describers network drawn with Networkx and Pyplot. Scotland is shown in blue, Wales in green and England in red.

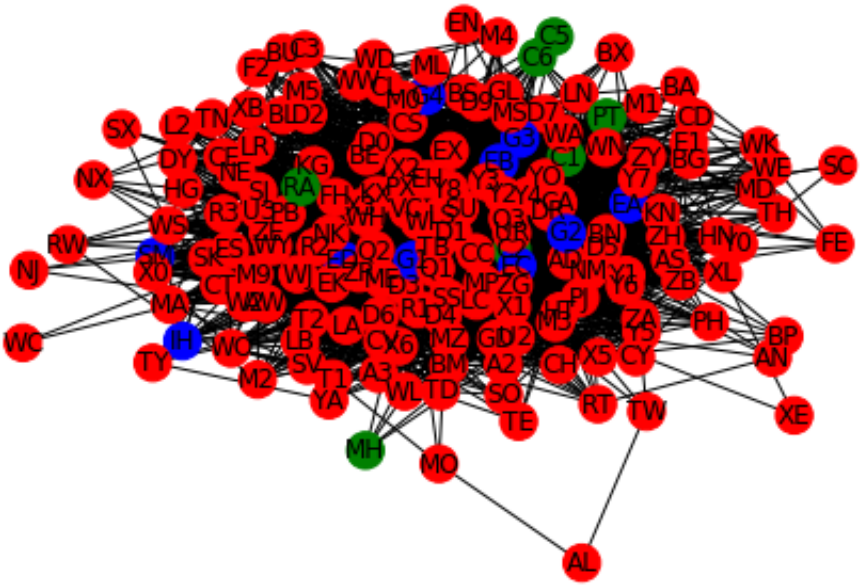

Figure 2.1: The Train Describer network plotted with pyplot.

In figure 2.1 we can see that Scotland is not accurately shown. This also applies to the Welsh Train Describers. The Inverness Train Describer is connected to Lancing on the south coast despite Inverness being the most northerly Train Describer in the dataset.

The use of some schedule data was mentioned in section 2.2. This schedule data will be used now to help solve the connection issues seen in figure 2.1. A longer breakdown of the schedule data can be seen in appendix A.

The schedule's inclusion of the headcode and the train UID allows for the removal of non-unique headcodes that are creating the large number of incorrect edges. On this particular day of data, the schedule had 49,560 different unique train UIDs and only 10,530 unique headcodes. This means that on average, each headcode refers to around five services running on the track. Of the unique headcodes, 2,758 are associated with only one train UID. This causes an issue because only headcode data is within the CA messages so different services are collected together and then are treated as the same service. This adds large amounts of false inter-Train Describer connections. If we now examine only the headcodes with one associated trainUID, the berth movement data will be far cleaner without multiple trains utilising the same headcode obscuring smooth movements by creating false connections.

Figure 2.2 shows what the connections between the describers would be like

with this step.

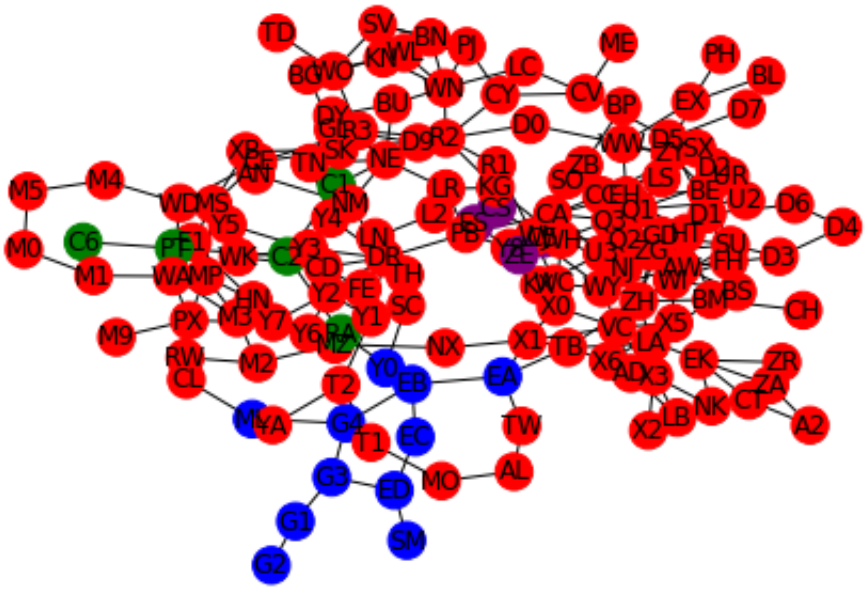

Figure 2.2: The Train Describer network plotted after the removal of duplicate headcodes.

Clearly the second diagram is more useful both visually and also for network analysis purposes. It features only 462 edges now, from 6334 edges in figure 2.1, a decrease of 93 percent.

Figure 2.2 consists of 152 nodes and 462 edges with an average degree of 3.04. Average degree is the mean number of other nodes that a node in the network is connected to. This is a sparse network with a density of just 0.0225, which is to be expected. Density is a measure of connectivity within a network. The value is between zero and one where zero is a network with no edges and one is a network where all nodes are connected to all other nodes. The real-life network spans Great Britain so it is unrealistic to expect some Train Describers to be adjacent to others. It consists of three connected components; one of which comprises of 148 nodes. The other two components are just two nodes connected by a single edge. These disconnected components are shown on figure 2.2 as purple. They are found in the centre of the diagram. The first of these components is ES-ZE two Train Describers which are now decommissioned and at the time of my data set ES (East Sussex) was entirely contained within ZE (Eastbourne). The other small component is the pair C5-CS which link Cheshire and North Wales which are linked on the real network.

The reason that this section remains unconnected is due to low traffic in the rural areas. This issue would be rectified if more data was analysed and a better method for handling incorrect attachments was devised.

The diameter of the largest component is twelve. There are a few paths of this length, one of which is Paisley (G2) to Plymouth (PH). The reason for this is obvious since Paisley is located near Glasgow and Plymouth is in Cornwall. The transitivity of the graph is 0.219 which is quite high when compared to the next network examined. This shows that high connectivity is valued on the network, allowing for more options of routes.

AreaID	Location	Betweenness Centrality
DR	Doncaster	0.18098
VC	London Victoria	0.17566
PB	Peterborough	0.15438
CA	Cambridge	0.13671
X1	London Bridge	0.13451

Table 2.2: Highest betweenness centrality for a node in the Train Describer network.

Table 2.2 displays a ranking of the Train Describers with the highest betweenness centrality. This measure of centrality was chosen because it finds hubs that contain important routes. It does this by counting the number of shortest paths through a node [32], [8]. The shortest path is a path between two nodes on the network traversing the fewest number of edges. After all the shortest paths are calculated the betweenness centrality can be found. The table shows that Doncaster is a hub for the network of Train Describers which makes sense because the station is a historically significant hub of the railway network. It is a major passenger interchange in south Yorkshire. The station with the second highest betweenness centrality is London Victoria which is a significant mainline rail station as well as a hub for other modes of transport such as bus and underground rail. However, Peterborough and Cambridge are not considered to be hubs and are likely more significant because of the number of berths in the Train Describer. Upon observation of figure 2.2, the Train Describer CA (Cambridge) is connected to many important describers such as King's Cross and is located near the centre of the graph. This is why it has such a high value despite Cambridge itself being a minor station on the network.

Area ID	Location	Eigenvector Centrality
ZG	Guildford TDC	0.33071
GD	Guildford	0.32839
EH	Eastleigh	0.31387
WI	Wimbledon	0.27865
Q2	Shenfield	0.27773

Table 2.3: Highest eigenvector centrality for a node in the Train Describer network.

In table 2.3 the eigenvector centrality is shown. This method makes use of the adjacency matrix. The adjacency matrix stores which nodes are connected with an edge, a one if connected and zero if not. Because the adjacency matrix is symmetric all of its eigenvalues are real and its eigenvectors are orthogonal and it is diagonalisable [7]. Eigenvector centrality uses the normalised eigenvector from the principal eigenvalue of the adjacency matrix and the entries in the eigenvector correspond to the nodes on the network. The largest values in the eigenvector have the highest eigenvector centrality and are considered the most central nodes in the graph [87]. The way to interpret a high value is the node is connected to many other nodes that have a high value [68]. The four Train Describers with the highest eigenvector centrality are all in the Wessex signalling area. They are connected to each other in the highly concentrated West London area. The last of the top five Train Describers, Shenfield, is also connected with the other four areas but is located in East London.

Area ID	Location	Page rank
DR	Doncaster	0.01896
VC	London Victoria	0.01425
EH	Eastleigh	0.01322
CA	Cambridge	0.01271
WI	Wimbledon	0.01267

Table 2.4: Highest page rank centrality for a node in the Train Describer network.

Table 2.4 shows the page rank centrality, a measure that is similar to eigen-

vector centrality, the major difference is achieved by normalising the values and introducing a random jump [104]. This algorithm was developed by Google founders Larry Page and Sergei Brin [23]. This was to develop a search algorithm for web pages. It increases the importance of a directional link based on the significance of the node the edge originated from [22]. Interestingly, all five Train Describers from the page rank comparison are in the two previous tables. This is not that surprising when compared to the eigenvector centrality since the page rank uses a similar algorithm but seeing some of the same describers from the betweenness table shows they are of particular interest.

Area ID	Location	Sum Weight
R3	Stafford	4853
SK	Stoke	3273
X1	London Bridge	3135
GD	Guildford	2542
ZG	Guildford TDC	2457

Table 2.5: Highest number of train observations at a Train Describer.

Another way of analysing networks is looking at node properties. Table 2.5 shows the weight of connections to a given node. Note that the number of observed movements in and out of a describer is what makes up the weight value. The connections between these regions are made by tracking individual headcodes and monitoring changes in the Train Describer section of the data. Note however that London bridge as well as the two Guildford Train Describers rate highly by this measure as well as some of the other centrality values.

This section has allowed us to examine the layout of Train Describers on the island of Great Britain. We have seen that the hubs on the network are made up of historically important locations like Eastleigh [11] and London Bridge. The work also shows that the Network Rail managed network still remains centred on London with most of the hubs found to be centrally located in the Greater London area. This could show that some areas of the network lack greater connectivity. Some of these smaller lines were removed or decommissioned during the Beeching cuts [20] It would have been interesting to see what these measures would have been if those

lines had remained.

The conclusion of this section is that the structure of the Train Describers network in Great Britain lacks geographical coverage of areas such as Wales. The Beeching cuts have hurt rural connectivity. London has been the focus of this railway as the centrality measures show. That is why we will look at Train Describers close to London in later chapters of this thesis. If more data regarding the geographic size of the areas could be obtained, then a better representation of connectivity could be calculated. Passenger numbers would be interesting and could be of greater value than the train observation numbers shown in table 2.5.

2.4 Visualising the full berth network model

This section of the chapter will be a discussion of the national-level berth network. The main problem to deal with in this sub-chapter is the scale of the network itself.

There are a substantial number of berths in the national network with approximately 28,000 observed on one day alone.

The way that trains go between areas is complex. As can be seen in the breakdown of the data in section 2.2, a berth movement can only happen within a single Train Describer. From this we can see that between describers there exist transition zones. These transition zones are areas reported by multiple regions but only one set of physical infrastructure underlying them. This means that the berths are shared and have a different berth ID for each overlapping describer. An example of this behaviour can be seen in figure 2.3.

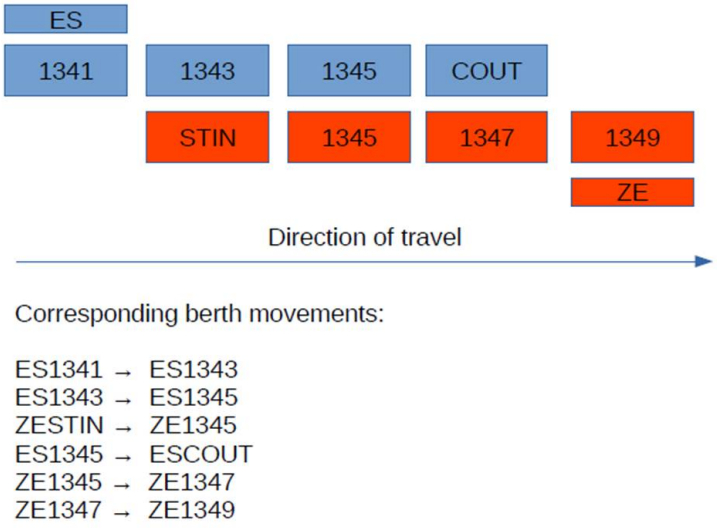

Figure 2.3: An example of a transition between two Train Describers.

In this diagram the blue berths are in the signalling area of East Sussex and the red berths are in Eastbourne. In this diagram the berth "1345" is being reported by both areas yet is only one physical berth. STIN is an abbreviated way of saying step in and COUT is short for clear out. These berth IDs are not themselves berths but are a representation of entering and leaving Train Describers. Note that this is only an example of an area transition and that there are other ways of moving between areas, using for example the CB and CC type messages that were mentioned in section 2.2. These CB and CC messages roughly translate to STIN and COUT but are instead stored in different message types. CB and CC messages are also used for removing trains from the network altogether.

There is no record of which berths are edge berths. This is because such information is known by the individual signalman administering the train describer. There is no one rule that defines a berth as an edge berth, instead these berths possess certain characteristics. We created four rules to detect these edge berths properties. These berths come in pairs, one in each Train Describer. The rules do not prove that a berth pair is an edge berth pair but the more of them that you satisfy the higher the likelihood is that they are. Note that these rules only apply if you are looking only at berth movements from a single headcode.

1. The two berth movements happen within one second of each other
2. The berth movements happen in adjacent areas
3. The berths have the same IDs

4. The berth movements follow a STIN and precede a COUT

These rules are used later in section 2.5.

In the same way as section 2.3 headcodes are separated and followed through the network to construct a graph. It matches up movements that occur between Train Describers and links them up by adjacency in time.

The next diagram (figure 2.4) shows an example of the large British railway network that is hard to visualise.

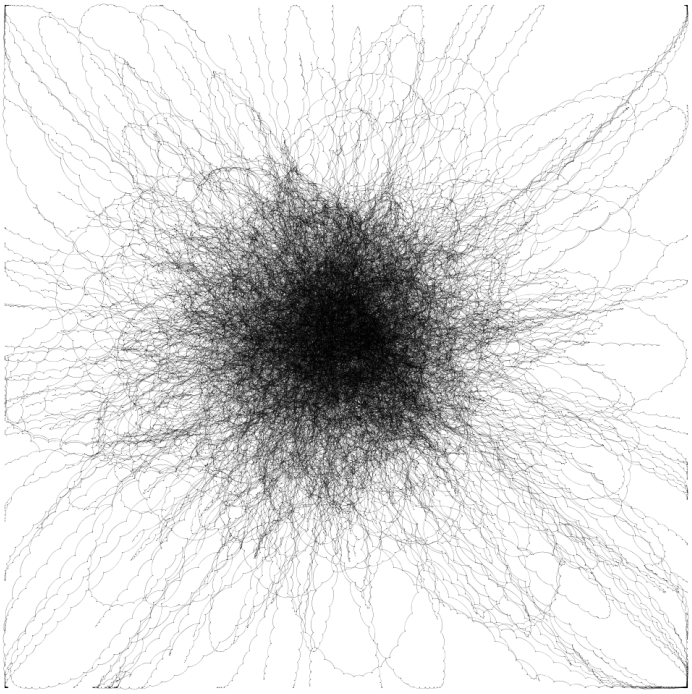

Figure 2.4: The entire British railway network drawn at a berth level.

This figure has no geographic information but has some interesting structures. The display algorithm Force Atlas 2 concentrates larger components towards the centre of the graph and pushes smaller unconnected components out. There are also large station components that exist outside of the heavily connected middle. This figure was included to demonstrate that berth visualisation can only be achieved at a local level. This is worked on in the next section.

It may not be possible to have an interpretable visualisation of this graph but network analysis can be applied to it. The network is significantly larger with 23,445 nodes and 33,644 edges having an average degree of 2.87. It has a much smaller

density value of 0.00012. This is not surprising since a lot of the berth network consists of long strings of continuous nodes that have one edge in and one edge out. The diameter of the largest component is 197, again this value is far larger than the area-level network.

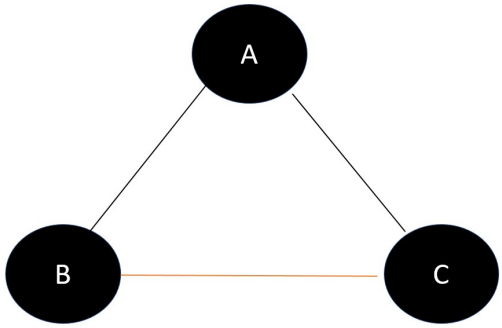

Figure 2.5: A diagram allowing for an easier explanation of transitivity.

Transitivity is given as a value between zero and one. It is a measure of how many triangles are found within a network compared with the maximum that could be formed. Figure 2.5 shows a graph with a transitivity of one since there is only one triangle that could be formed, and the triangle is shown here complete. If the edge from node b to node c were removed, then the triangle would be broken. This results in the transitivity reducing to zero. The transitivity of the British berth network shown in figure 2.4 is 0.0277 which is low as expected with the low-density value.

Smaller sections of this network shown in figure 2.4 can be represented more easily and some of the smaller Train Describers can be shown in their entirety without losing all visual clarity. This network consists of 73 disconnected components one of these is the main component with more than 23,000 nodes and the other 72 are disconnected pairs.

The three most important berths by betweenness centrality are all in the Leamington Corridor Train Describer. This Train Describer covers a line of track next to Royal Leamington Spa so is in the Midlands. The fourth berth on this list is also close to the Leamington corridor, at least geographically. The last berth is in the York Train Describer which is a hub in the North-East of England.

BerthID	Location	Betweenness Centrality
LC7302	Leamington Corridor	0.21653
LC1293	Leamington Corridor	0.14106
LCSB01	Leamington Corridor	0.13169
R14126	Rugby (Watford-Bletchley)	0.12057
Y10227	York	0.10942

Table 2.6: Highest betweenness centrality nodes of the berth-level network.

BerthID	Location	Eigenvector Centrality
HT0314	Havant	0.26598
ZGH314	Guildsford TDC	0.26591
ZGH312	Guildsford TDC	0.25565
ZGH316	Guildsford TDC	0.25335
GDH314	Guildsford	0.23710

Table 2.7: Highest eigenvector centrality nodes of the berth-level network.

The highest rated berths by eigenvector centrality have some more recognisable Train Describers with Guildsford and Guildsford TDC. Interestingly, if you look at Figure 2.1, as well as the edge berth rules it seems likely that ZGH314 and GDH314 represent the same physical berth.

BerthID	Location	Page rank
LC7302	Leamington Corridor	0.00025
LC1293	Leamington Corridor	0.00020
NX0266	East London Line	0.00017
ME0388	Marylebone Chilterns	0.00015
CT018C	Channel Tunnel	0.00015

Table 2.8: Highest page rank centrality nodes of the berth-level network.

The page rank algorithm shows the same berths as the betweenness centrality ranking highly.

The communities analysis shows individual berths normally fall into groups similar to their respective Train Describers. This is not surprising since all inter-area connections are added through observation. This leads to weak connections between areas. They also can partner together with adjacent areas to form larger communities.

2.5 Visualising the berth-level network at a local level

2.5.1 Introduction to the Fenchurch St to Shoeburyness line

Section 2.4 showed the whole network, but this section will focus on just a few Train Describers. It will also present solutions to local issues that, whilst present in the global data set, are more noticeable at a smaller scale. Our choice for a local analysis is the Fenchurch St to Shoeburyness line. It was chosen due to low levels of freight trains as well as great berth coverage and a single passenger operator, C2C (Coast to coast). The single operator is important because operators compete when using the same track sometimes causing timetabling issues. The objective of this section of work is to address the discrepancy between digital berths and the underlying physical berths. This will result in an accurate representation of the underlying physical network as observed from the berth movements. This will be done by removing edges that do not exist or by examining nodes that may only exist digitally.

Figure 2.6: Line as illustrated by the operator C2C[13].

The important features of the line can be seen in figure 2.6 [13]. The Tilbury loop branch line is at the bottom edge of the diagram, next to the river Thames. The Liverpool St line is at the top of the diagram connecting only at Barking station. Barking station is the largest and most complex structure on the line from an operational point of view. Below, in figure 2.7, can be seen an early attempt of ours to show this line at a berth-level of detail. Note that since this line is bi-directional there are two of each station shown, eastbound and westbound, if that station has no method of travelling between the two sides of track.

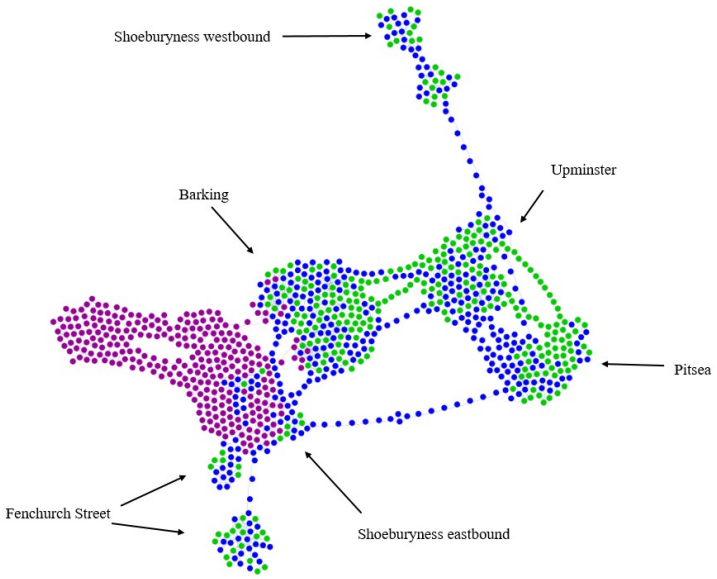

Figure 2.7: Connected berth diagram of the Fenchurch St to Shoeburyness line drawn in Gephi.

2.5.2 Snooping berths

In figure 2.7 the three colours show different Train Describers. Blue is the UR Train Describer which makes up most of the mainline, on figure 2.6 this is the section between Fenchurch St on the left of the diagram and Shoeburyness on the right. Green shows the U2 area that covers a branch line to the south of the mainline on figure 2.6. Purple is Q1 which is the line to Stratford, shown as the dashed line above West Ham in figure 2.6. There are two different Fenchurch St and Shoeburyness' because they are shown in both directions, so they represent Fenchurch St Westbound and Fenchurch St Eastbound. There do exist connections between the two sides of the track, but they are not used in normal service so are not shown on this diagram. Trains do also turn around at these locations but the service ends when the train reaches its destination and a new service begins which travels in the other direction. Since the only metrics that can be used for train classification are headcode and trainUID and they change when a service terminates it is not possible to track these turnarounds accurately.

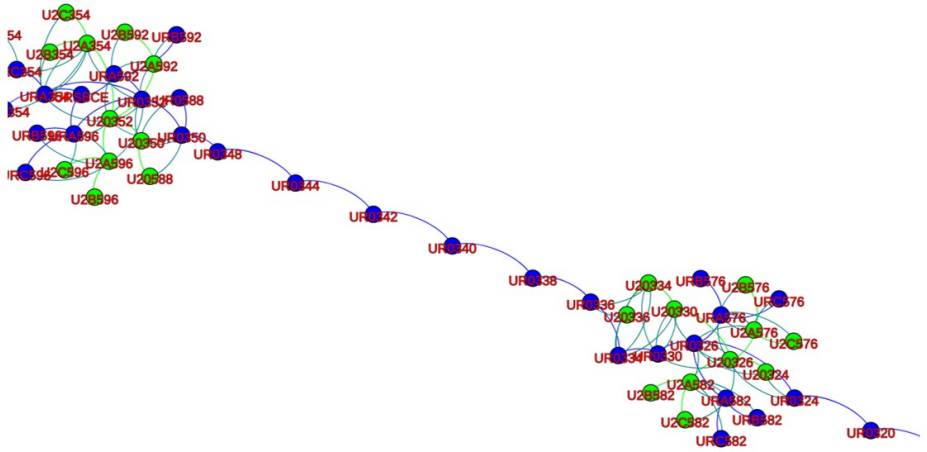

Figure 2.8: Magnified view of Shoeburyness westbound.

Figure 2.8 shows a magnified view of Shoeburyness westbound and Thorpe Bay as defined by their berths. There are two different signalling areas, UR and U2, mixed together the latter of which is out of place. This situation is caused by the signalman from the U2 region observing the UR section of the network.The U2 region is shown by the green berths in this case. This interest in other regions is behaviour shown by signalmen on a national level. These operators were historically only interested in regions under their control. Over time they desired to know what was happening near to their area of influence so they could prepare for incoming disruptive trains. The way that this was implemented was to allow berths to be in multiple Train Describers creating a phenomenon we call "snooping berths". This means that a Train Describer can consist of many disconnected components some of which are short snapshots of other areas.

This behaviour leads to a more complicated network layout that does not represent the physical infrastructure. Therefore, the next task is to remove the duplicate berths which will allow us to see the underlying topology.

Snooping berths and edge berths (mentioned in section 2.4) have very similar behaviour since both copy movements from a single physical berth. However, there is a way to differentiate them at a local level. The method works as follows:

1. Compare two or more Train Describers but only where the berths have the same identifiers

2. Verify that at least 80 percent of the movements have a match within one second, if not, stop comparing this berth pair

3. Check if the the berths are part of the largest connected component in their respective Train Describers

4. If both berths are in the largest connected component then this is an edge berth, if only one of the berths are in their respective largest component then that berth is the original and the other berth is a snooping berth

This method characterises almost all the duplicate berths with 96.3 percent of berths receiving a classification. The remaining berths are duplicate berths but neither one of the berth pair are in the largest connected component of their respective areas.

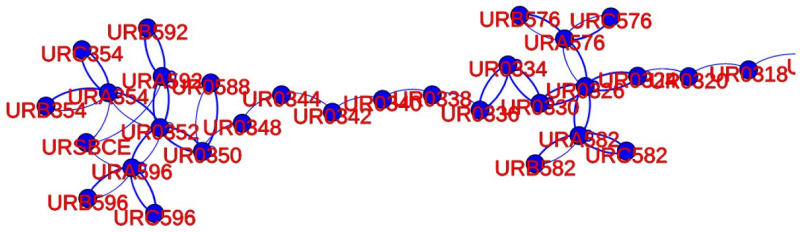

Figure 2.9: Same set of berths as in figure 2.8 but with the Snooping berths removed.

2.5.3 Edge Berths

Unfortunately, there are examples of three Train Describers that overlap. One such case is at Barking station where the areas UR, U2 and Q1 overlap in our dataset. This area contains a berth that best symbolises the duplicate berths problem where a single physical berth reports as six different berth IDs. This is only possible since the berth is considered to be in three Train Describers and is also snooped on by all three areas.

By implementing the above algorithm, we removed the majority of snooping berths and also combined the edge berth pairs into single berths. These singular berths were then placed into a new dummy Train Describer which is shown in yellow and called “UR”.

The below figures show a comparison of Upminster station which displays the difference made by fixing duplicate berths.

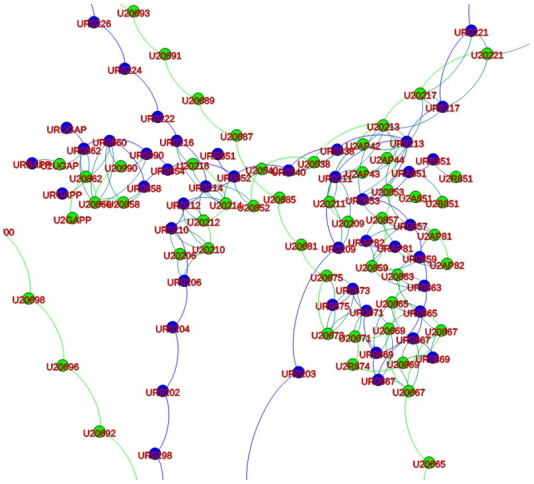

Figure 2.10: Magnified view of the berths comprising Upminster station.

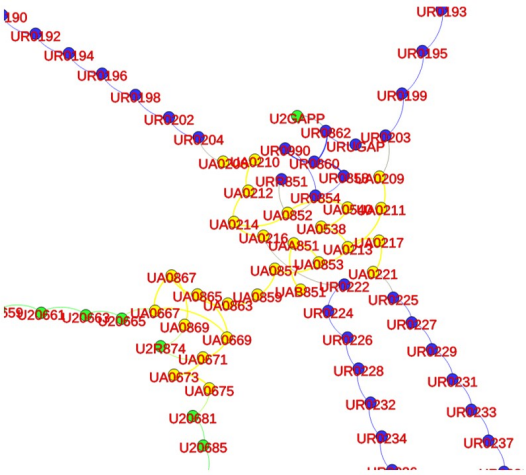

Figure 2.11: Upminster Station with edge berths merged into single berths.

Upon observation of the berth identifiers in these pictures we can see a reduction in the number of berths. The number of berths has decreased from 930 to 749, a reduction of 19 percent across the line. This is due to the removal of both the snooping berths and the combining of the edge berths. The new yellow Train Describer creates a buffer zone between the blue and green berths, this is the expected behaviour that was anticipated. Upminster should be a station that connects the mainline to the Tilbury loop. This connection is simplified with the combination of

the edge berth pairs. The network now resembles the physical berths more closely.

2.5.4 Duplicate headcodes

There are still problems with connectivity. Many of the issues and inconsistencies left to solve are caused by duplicate headcodes. In section 2.3 we used a simple fix to remove these trains entirely. For this local section we require a more precise solution.This is due to the way that trains are named. The table below shows a breakdown of train headcodes:

Train type	Destination/Route	Train number
1	G	29

Table 2.9: Breakdown of typical Headcode for a service.

The train number shows the quantity of trains with the same first two digits seen on that day in that Train Describer. So, the above example would be the 29th "1G" train of the day in that Describer. The train type shows the classification of train, type one trains are express passenger services, type two trains are local passenger services and so on. The Destination/Route column does not express enough variety for the whole network, so each letter varies in meaning around the country. This is why this headcode lacks uniqueness.

The rules we devised for detecting suspicious headcodes are as follows:

1. Extract the berth movements for a single headcode.
2. For each berth movement check if the Train Describers in the previous and subsequent movement have different IDs.
3. If both differ check if one of the movements either side happened within one second.
4. If neither happened within the allotted period flag the movement as suspicious.
5. If a headcode exceeds five suspicious movements the headcode is likely a duplicate headcode.

These parameters are designed to flag berth movements that are not legitimate. Step two flags a number of inter-area movements and step three uses a rule similar to that of the edge berths classification to allow for legitimate inter-area movements which is why one second is used as a buffer. Step five specifies five

suspicious movements as the limit because a border region near Barking station repeatably flagged up to three suspicious movements for every train that travelled down that section of line. Removing all these trains would leave gaps in the topology and these trains did not cause any incorrect links.

The above method locates disruptive duplicate headcodes. These are two separate trains that are moving in this local section of the network at the same time and therefore cause the highest number of falsified connections. Other duplicate headcodes do not pose an issue because one of the conditions required for an inter-area movement (a movement between train describers) is that the concurrent movements do not happen more than twenty minutes apart.

Figure 2.12 shows the disruptive effect of these duplicate headcodes.

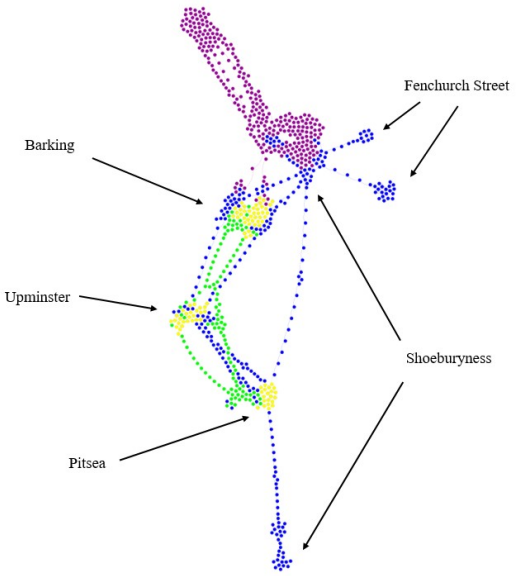

Figure 2.12: Connected berth diagram showing the effect of duplicate headcodes.

In the above figure, Shoeburyness is incorrectly attached to the train describer Q1. These movements are created by two trains with the same headcode running at the same time, in this case it is a train called "2F30". This train has been flagged using the above algorithm so can be removed. Figure 2.13 shows the network without this train and now correctly displays the network in this area. The number of edges removed with this process is fifteen. Overall, the removal of snooping berths as well as edge berths and now duplicate headcodes results in 415 less edges. This network now shows the physical berths that make up the real network. Berth diagrams are available within the industry for individual Train Describers, but they

do not attempt to show the topology of multiple Train Describers combined.

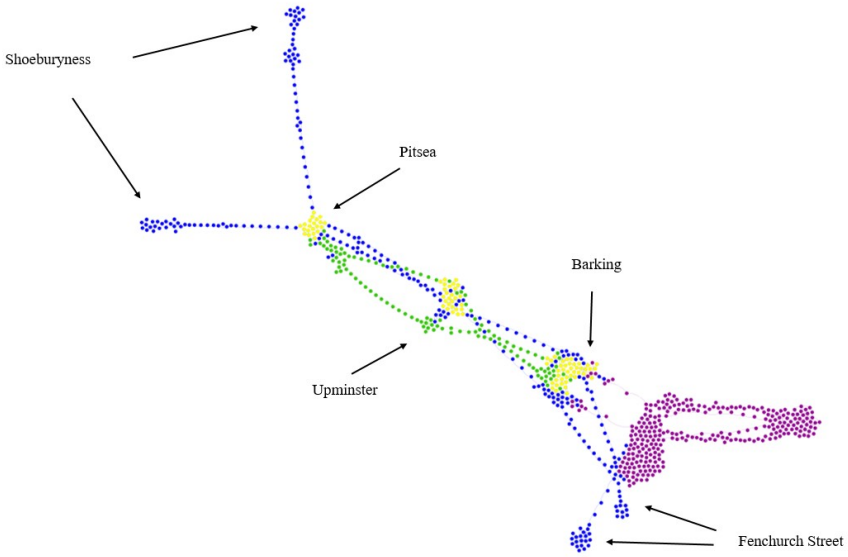

Figure 2.13: Connected berth diagram of the Fenchurch St to Shoeburyness line drawn in Gephi.

Other areas of the network have also been looked at with this procedure. the Train Describers for East Sussex and Eastbourne were examined for understanding the extent of snooping berths. All of the berths in the East Sussex Train Describer were snooped upon by Eastbourne. Extending this procedure for the whole network would require a large amount of future work. The method of eliminating duplicate headcodes only works with local problems so only a few Train Describers can be simplified at a time.

2.6 Conclusion

This chapter has conveyed the difficulty of using berth data to reverse engineer the railway network. We devised algorithms to counteract this complexity and to clean the data set. This process allowed us to perform meaningful analysis on this output. Subsequently, we worked on analysis of different scales of the network and worked on creating the digital model from the data to mirror the physical one.

To illustrate our earlier point from the introduction of this chapter, regarding the changing nature of the network, my data set is from September to October 2019 and already many of the Train Describers have now been decommissioned. Due to

the continued projects of improvements and re-signalling the network will continue to evolve and change all the time. It is likely that some of the currently analogue areas will be digitised, adding completely novel Train Describers to the network. This means that this will always be a work in progress where the construction of the digital model will be behind the physical until all of the track is in the same upgraded state. Currently the U.K. government is spending in excess of 100 billion on creating a new high-speed rail line HS2 [9]. After the partial cancellation of HS2, funding has now been granted for upgrades to existing railway infrastructure[93]. HS2 is managed by HS2 limited, a wholly owned government subsidiary [94]. Just like Eurostar (HS1) it will not be included in the Network Rail managed railway. Importantly, the data it produces will not be published on Network Rail's datafeeds because it will not be run by Network Rail. Now funding has been granted for rural lines they will be re-signalled and they will need to be viewed with new data. This will reduce travel times and hopefully lead to a reduction in the emphasis on London in the overall network structure.

Visualisation of the entire Great Britain network is possible to do with a display algorithm but unfortunately it is hard to glean any useful information from the picture. Drawing the berth network on a small scale does allow for worthwhile analysis though, as shown in figure 2.13. If we had access to more geographical information such as GPS locations of trains this would better allow placement of berths in the diagrams. There is such a plan to integrate GNSS (the European equivalent of GPS) into the Darwin system [25], [54] that currently tracks trains around the network and displays information to passengers at stations.

The network analysis shows that the hub of the global network is found just west of London. This is not surprising since all of the top four busiest stations in Great Britain are in London [85]. Further analysis could be done on different types of community analysis. The berth-level network by its nature will be sorted into communities resembling Train Describers.

The problem of duplicate berths is an unnecessary one caused by the lack of two entries in the data for Train Describers. If such an action was taken, then border regions could be removed altogether since inter-area movements would no longer need to be inferred and they could instead be observed. They also show the local rather than global perspective that is pervasive on the rail network in Great Britain. It is likely in the future that there will be a greater number of snooping berths as signalmen grow more interested in seeing further afield. Hopefully the logical end of this is to change the data set and allow inter-area movements into the data set.

Duplicate headcodes are of much more concern since our solution was simply to filter out certain problematic headcodes. This approach can work locally but cannot be applied at a global level. This problem could be better fixed by Network Rail by either adapting the data set to have a more unique identifier or by introducing new policies banning adjacent Train Describers from using the same headcodes.

Chapter 3

Analysis of a Single Line's Infrastructure

3.1 Introduction to the Fenchurch St to Shoeburyness line

A railway line has many problems to contend with on a daily basis but many of the issues experienced are caused by the underlying infrastructure. In this body of work, we will look at various aspects of infrastructure analysis including bottlenecks, where the flow of traffic on the network is at its most constricted. Also examined here will be the flow of trains in both directions to see how delay changes over time.

This chapter will be focusing on the Fenchurch St to Shoeburyness line, as seen in figure 2.5. Most of the analysis this time will focus on the mainline largely disregarding the branch line known as the Tilbury loop. The main line is primarily a commuter service, characterised by high passenger traffic westbound in the morning and the majority of passenger traffic travelling in the other direction in the evening. It also features more services at peak times during to allow for higher passenger flow.

In general, the closer the line is to Fenchurch St the greater the number of services there are on the line. This is due to the rejoining of the branch line and additional trains coming from the Liverpool St section that comes from North London at Barking.

Commuter lines are generally run at a high capacity, but capacity is always part of a balance with resilience to disruption. This is because trains that are nearer to each other have less leeway to cope with delays. To examine this line, we can better access specific data for trains. This data is contained within the timetable

data set. Unfortunately, the timetable data received required significant work to be used and also needs to be related to the berth data described in the last chapter.

Section 3.2 and appendix A describe the different data sets and how they are processed for usage. There are four data sets used here. The first is the schedule data itself and the other three are intermediate sets of data that are used to translate different aspects of the data set to allow for accurate combining.

3.2 Data Linkage

This section will discuss the linking of the data sets through shared fields. The diagram below shows how the schedule data was connected to the berth movements data.

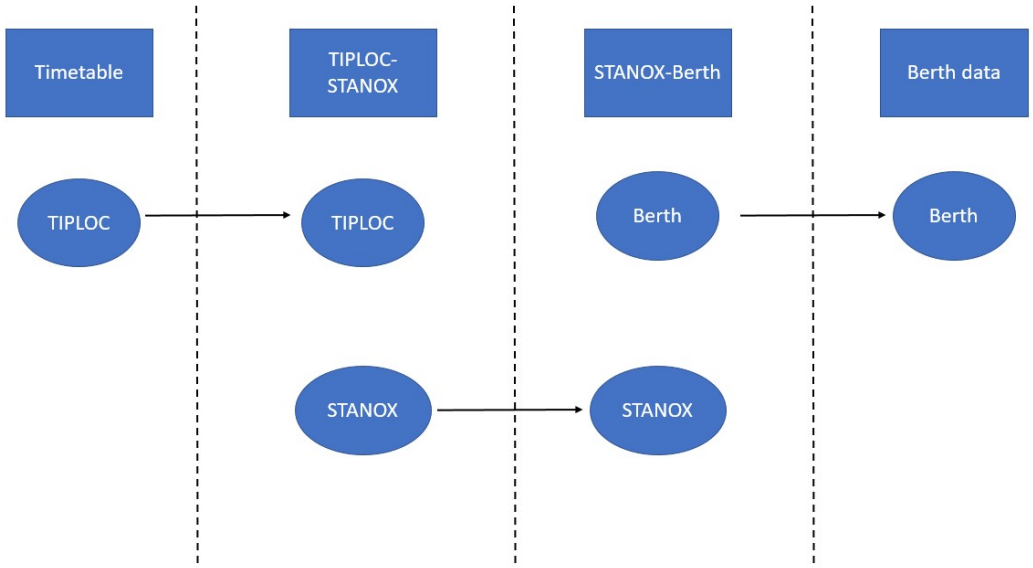

Figure 3.1: Visualising the links between the different datasets used for this thesis.

The first step was to combine the useful information from the TIPLOC-STANOX data set with that from the STANOX-Berth data. This effectively created a TIPLOC-Berth data set, this is shown below.

Train Describer	Step Type	From Berth	To Berth	STANOX	Platform	Passing Offset	Stopping Offset	TIPLOC
IH	I	NaN	7904	01125	1	0	0	NAIRN
IH	I	NaN	7906	01125	2	0	0	NAIRN
IH	I	7903	7909	01125	1	0	0	NAIRN
IH	B	7904	7909	01125	1	25	25	NAIRN

Figure 3.2: Sample of the STANOX-berth dataset.

This was then sorted to remove all step types other the "B" types since they are the standard type of movements. The other types of movements represent abnormal behaviour such as adding or removing a train from the network. The next step was to append the trigger berths and corresponding offsets onto the timetable data. This then gives the schedule access to the proposed berths that could correspond to arrivals at the stated TIPLOCs. There are a number of ways that a train can be recorded arriving at a TIPLOC therefore this mapping is a one-to-many relationship.

The next step is to append the berth movements with the relevant schedule information. This leads to the sample of data shown in figure 3.3.

Area	From	To	Headcode	UTC time	TIPLOC	Scheduled arrival time	Offset	Adjusted arrival time	Delay
UR	0344	0342	2D01	2019-10-01 04:06:45	THOPBAY	2019-10-01 04:06:00	-41	2019-10-01 04:06:04	00:00:04
UR	0342	0340	2D01	2019-10-01 04:07:16	NaN	2019-10-01 04:07:16	0	2019-10-01 04:07:16	00:00:00
UR	0340	0338	2D01	2019-10-01 04:07:54	STHNDE	2019-10-01 04:08:00	0	2019-10-01 04:08:29	00:00:29
UR	0338	0336	2D01	2019-10-01 04:0920	STHNDE	2019-10-01 04:08:00	25	2019-10-01 04:06:50	00:00:50

Figure 3.3: Breakdown of the appended berth data.

This shows the berth movements from chapter two with attached schedule data. It contains all the fields from the berth data, the only difference is that the time is now in a date-time format more useful for analysis. It now features a TIPLOC field which is occupied if the berth movement corresponds to a valid location. It also has a scheduled arrival time from the timetable. This field is a copy of the berth reporting time if there is no valid TIPLOC.

If there is a TIPLOC attached there is also an offset and this offset is used to adjust the berth movements to the platforms for comparison to the scheduled arrival time. The last column shows the delay currently experienced by that particular service at that time. A positive number indicates lateness. The third and fourth rows show that the offsets are not perfectly accurate because otherwise the adjusted arrival times would be closer together. After this data linkage the data set is now ready for analysis.

3.3 SRTs for individual berths

3.3.1 SRT calculation

This section focuses on SRTs (Sectional running time). SRTs are the time it takes for a train to traverse a section of track. They are calculated by ordering berth movements by headcode and by time. Then adjacent berth movement timings are compared which then returns the time taken to traverse between berths. The berth movements are then given a unique identifier which allows data analysis for all individual berth pairs.

Area	From	To	ID	Median	Mean	STD	STD/ Mean
UR	0105	0107	UR01050107	00:00:56	00:00:56	00:00:11	0.200
UR	0106	R101	UR0106R101	00:00:19	00:00:34	00:00:34	0.995
UR	0106	R103	UR0106R103	00:00:17	00:00:23	00:00:24	1.022
UR	0106	R505	UR0106R505	00:00:18	00:00:25	00:00:26	1.072

Figure 3.4: A sample of Sectional Running Times.

Figure 3.4 above shows a sample of the SRT data created by this process which is calculated with five weeks of data. This data adds up to approximately 870,000 berth movements for the UR train describer alone. The unique identifier is created by combining both berth IDs with the train describer they belong to. This data can then be analysed to return the median, mean and standard deviation.

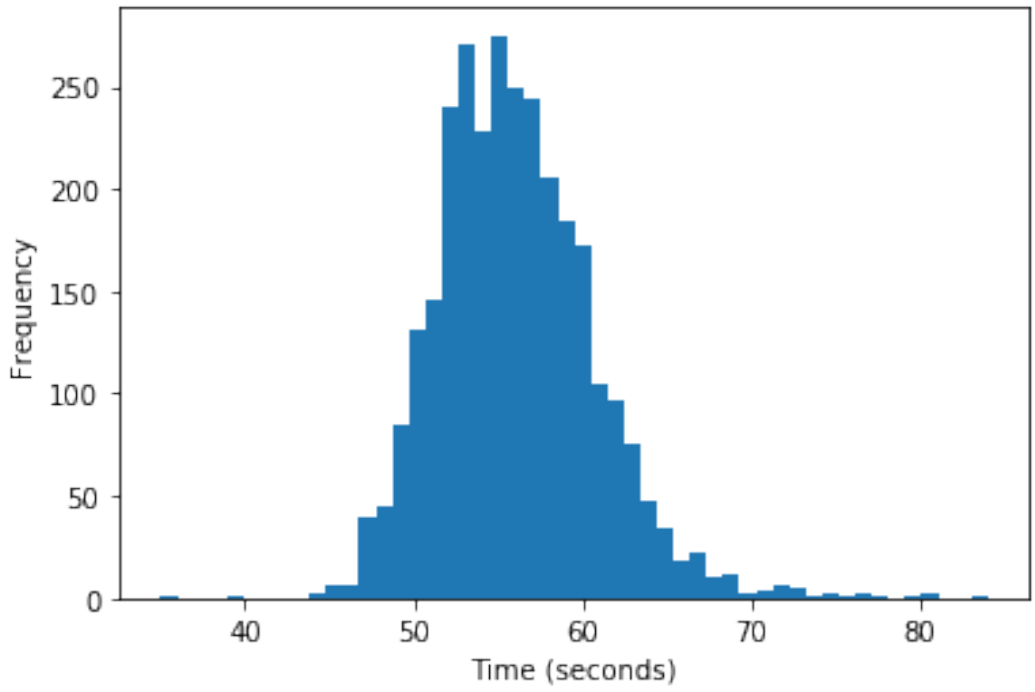

Figure 3.5: An example histogram of an SRT.

One interesting observation from the table shown in figure 3.4 is that the mean is always larger than the median. This reflects the underlying distribution which has a heavy tail. This is because sections of track have speed limits so trains will not traverse a section of track faster than a certain speed. However, there is no upper bound on the time it can take to traverse a berth. So, the typical travel time for a berth is close to the minimum leading to this distribution. This graph does have an outlier that was removed which took 600 seconds to traverse the track. These SRTs will be used extensively later in chapter four.

There is no measure of berth track segment length available. The physical infrastructure can be measured but there is no dataset that exists documenting berth length. Crossrail is an outlier because it is recording the location of all installed infrastructure [20]. If there were a list of maximum speeds allowed over berths, then the length could be found by dividing the speed by the SRT. In the future an accurate GPS system could be used to estimate berth length. The metric we used to examine berth length is the sectional running time.

The berths can be divided between station berths and non-station berths using SRT as the units.

3.3.2 Capacity modelling

The SRTs can be used to calculate capacity for section of line. We extracted a list of berths that are on the mainline from Fenchurch St to Shoeburyness by tracking a service from one end of the line to the other. This subset of berths can now be analysed to examine capacity.

As mentioned previously, berths are a safety system that is mainly used to maintain distance between trains. They ensure trains do not come less than two berths apart. Therefore, if we add the SRTs for two adjacent berths together we have the amount of time that trains will be apart from one another. Then if we divide one hour by the combined SRTs then we have a measure of capacity in trains per hour.

TD	Berth1	Berth2	Berth 3	Combined SRT	Capacity (trains per hour)
UR	0352	0350	0348	00:03:32	16.98
UR	0154	0150	0148	00:03:09	19.05
UR	0150	0148	0146	00:02:52	20.93
UR	0350	0348	0344	00:02:48	21.43

Figure 3.6: Three berth SRT capacity table with the largest combined SRTs.

The maximum value for combined SRT in figure 3.6 is 212 seconds in the westbound direction. This translates to 16.98 trains per hour as the minimum capacity for the line in the westbound direction. These berths (figure 3.6) are the respective pairs for sidings. The berths 0350 and 0348 are the pair for the siding near Shoeburyness and the berths 0150 and 0148 denote the East Ham depot between West Ham and Barking. They are unsurprisingly slower as the trains accelerate onto the line.

3.4 Bottleneck analysis of the London Fenchurch St to Shoeburyness line

The individual TIPLOCs on the line are useful to determine where a train is currently and they can be used to track a train's punctuality along the line. This section will examine these TIPLOCs to observe which experiences the most delay and what sections of the line increase the lateness of trains that pass through it. Part of this will be examining the line in both directions.

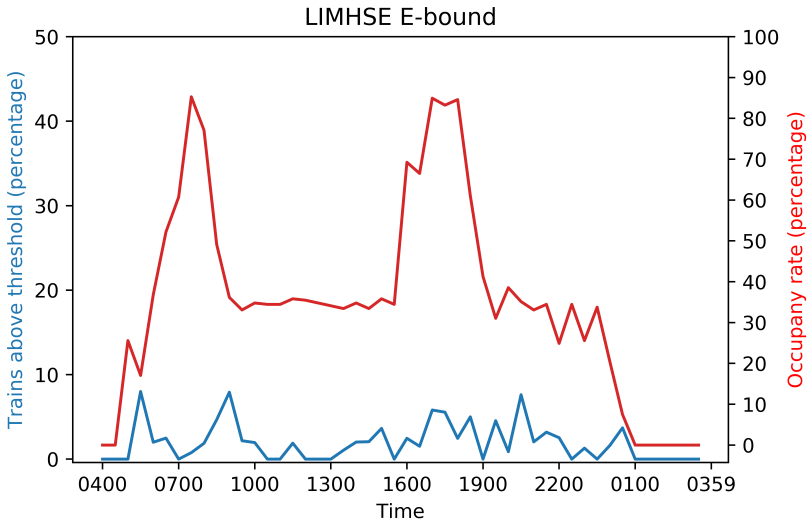

Figure 3.7: Plot showing low levels of super-threshold delay against the schedule at Limehouse.

Figure 3.7 shows Limehouse eastbound. The left y-axis is the percentage of trains arriving at this TIPLOC in this direction above the threshold. This is derived from the appended berth movements created in section 3.2. In this case the threshold is three minutes. This is the limit used by the Great Britain Delay Attribution Board (DAB) to decide whether a delay is worth investigating so it can be attributed to a Train operating company or Network Rail [62]. The threshold itself seems arbitrary but such delays are not investigated here and we will work on them more in chapter four. This approach to using thresholds instead of average delay was because average delay statistics are heavily influenced by outliers. When the average delay method is used the resulting graphs lack interpretability.

The right y-axis shows the occupancy rate which is the number of trains

arriving at the given TIPLOC in a half hour period as a percentage of the maximum capacity on this section of the line. We used the capacity measurements from section 3.3 but changed to using four-berth capacity because it is a fairer measurement for free-running trains during operation.

As can be seen from the figure, super-threshold delay at Limehouse Eastbound is very low. This is not a surprise because it is the first TIPLOC after departure from the first station on the line.

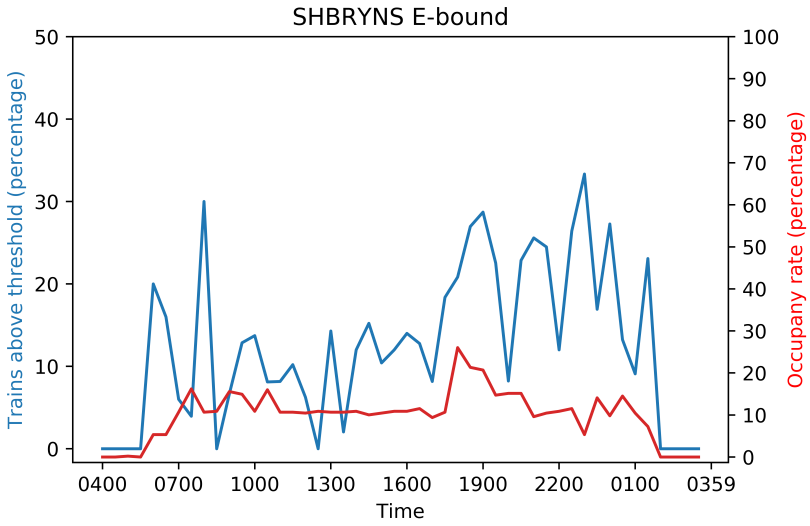

Figure 3.8: Plot showing high levels of super-threshold delay at Shoeburyness.

Figure 3.8 shows Shoeburyness Eastbound so these services are arriving at the end of their paths along the line. It shows significant super-threshold delay at times exceeding 30 percent of trains arriving whilst being significantly late. However, there are substantial fluctuations in the train delays. Initially, we thought that this may be caused by the timetable. That could show periodic arrivals of express trains that come only once per hour. But when the period was changed away from half hour sections this spiking behaviour did not go away. Of note is the turn around time. As shown below in figure 3.9.

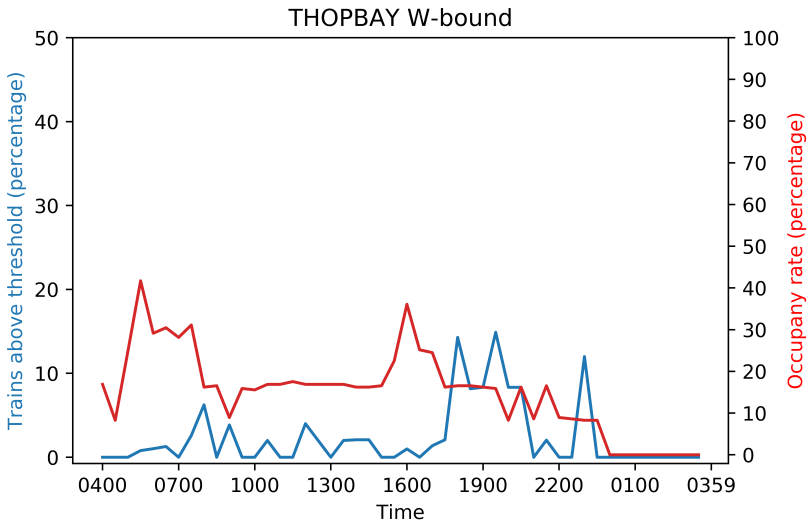

Figure 3.9: Plot of Thorpe Bay showing higher then expected levels of super-threshold delay.

Thorpe Bay is the first TIPLOC after Shoeburyness for the departure in the westbound direction. It would be expected that little to no super-threshold delay will be present. Despite this, the figure shows delay is present. We propose that this is due to the turn-around time left at Shoeburyness. The trains that depart from the station are the same ones that entered earlier causing departure delay due to insufficient leeway in the schedule. This causes trains to depart late and prevents full recovery in post-peak times. This in turn causes greater delays in the evening rush hour. Another observation that can be made from the graphs is that the line is much more congested between Fenchurch St and Upminster.

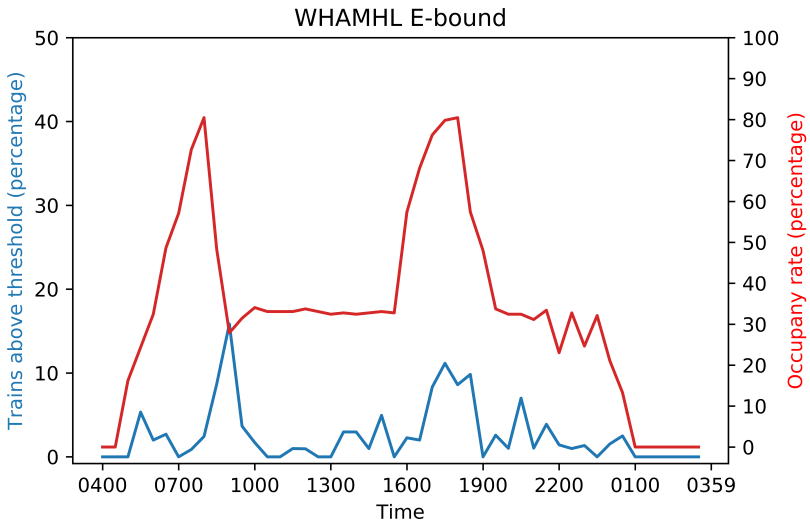

Figure 3.10: Plot of West Ham showing high occupancy rates at peak time of the day.

Figure 3.10 shows this high capacity and 3.11 shows a more ordinary occupancy rate seen on the line.

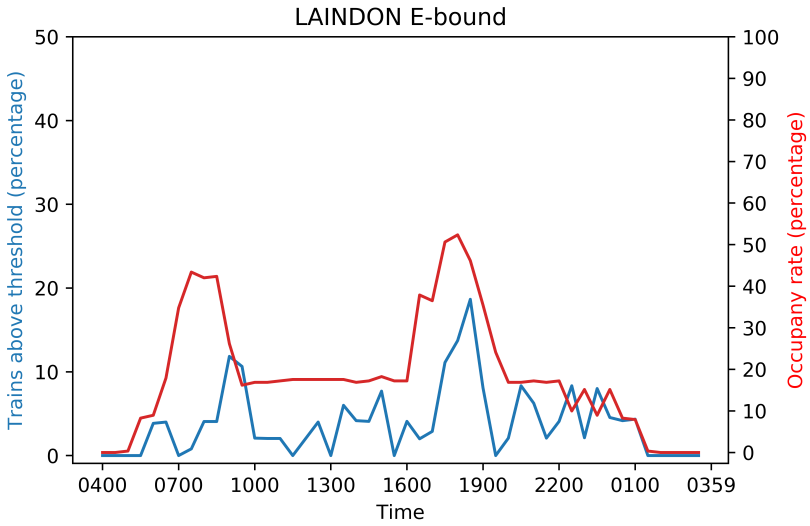

Figure 3.11: Plot of Laindon showing lower occupancy rates at peak times.

This higher occupancy is due to Upminster being one of the stations where services can depart to the branch line (Train Describer U2). But Barking is also influential here as it is the origin station for many short services to Fenchurch Street.

3.5 Attributing delay using a random forest regression approach

This section will focus on a random forest regression model that will allocate which features of a service are responsible for how late it arrives at any given TIPLOC.

To discuss random forest regression, we must first understand decision trees. Decision trees are a mathematical method that can be used for classification or regression [19]. It originates from machine learning [19]. It is also a method that is intuitive to interpret. This is due to the tree that can be examined during the analysis stage. Decision trees can also be used in combination with other methods as well. The method breaks down a larger question into a series of smaller ones that are trivial to answer. A disadvantage of the method is that as the tree grows accuracy comes at the cost of computation time [1]. Classification with decision trees works by partitioning the given dataset repeatedly until a sufficient class has been located. It can be used in many areas such as medical diagnoses or image classification [5].

The application that will be looked at here is the area of railway delay prediction. Below is an example of what a decision tree would look like in this area:

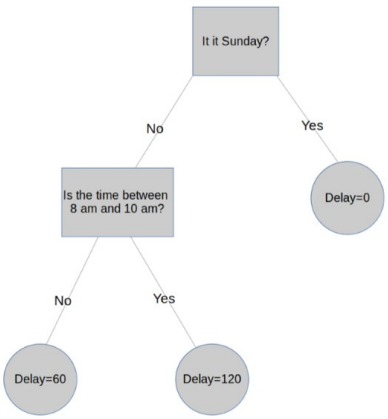

Figure 3.12: An example of a simple decision tree.

The purpose of the decision tree in figure 3.12 is to predict train delay based on submitted parameters. The decisions are expressed here as questions. The first is a check of the day that the prediction occurs on. Sundays typically have less delay than other days of the week. The second examines the rush hour that happens during the weekdays. The interpretability of this method can be seen in the decisions and the outcomes resulting from them. Outcomes can be traced back up the decision tree to find out how those results were decided. Decision trees are composed of two nodal types, branches and leaves. The decision nodes are branches and the results are the leaves here. The objective value of this tree is to predict delay against the timetable. The method will be evaluated by absolute median error when compared to the observed value.

With a sufficient sample size, the data can be split between training and validation, so that the tree can be evaluated for accuracy. Decision trees here will be used for random forest regression. Random forest regression requires an ensemble of decision trees [80] that are trained with data points and are then used to predict based on newly observed data. This is done via a process called bootstrapping, a method that reduces variance and overfitting. Bootstrapping works by sampling the dataset with replacement [59], this means that some samples are used multiple times across different generated decision trees. This method ensures independence between trees. An individual bootstrap dataset has duplicate data points to create an overall set the same size as the original. Each data point has differentiating

features which are then examined as to whether they provide a positive or negative effect on the overall prediction result. If we refer to the previous figure this step would be examining whether operating on a Sunday is positive or negative for delay. This feature is examined at all data points and combined to output a confusion matrix for each feature. This confusion matrix consists of predictions versus actuality. It tests the false positives as well as the true positives and the true and false negatives when used for predicting the desired metric [17], in our case train delay. Metrics can be calculated from the finished confusion matrices to rank the features by their value in prediction. The features can then be combined to create a decision tree that uses the important features to divide the dataset. So, if like the above figure, Sundays are found to be an indicator of low delay, and the given data point is observed on a Sunday the algorithm will proceed down the right branch. The is repeated for different features to create a decision tree of a desired depth [17]. When used for random forest regression the trees have an additional constraint that requires that each tree only utilises a random subset of the features this is known as the random subspace method [43].

The method is summarised in Ho's 1995 paper [42] as:

1. For a given feature space of m dimensions there are 2^m subspaces in which a decision tree can be constructed.

2. Randomly select a subset of features and use these to create independent subspaces.

3. For each of these subspaces of unique features construct a decision tree using all of the training data.

This ensures greater variability between the trees as well as independence. Later on, the method was developed to use bagging as well [86]. This is now a random forest. When used for classification a majority vote is used to decide the result but for regression a mean of the predictions is taken.

Originally the random forest method was described as a black-box method [79] due to its lack of interpretability and the seemingly unexplainable success. The lack of explanation can be a significant barrier when used for applications such as the medical field where life changing decisions require justification. This can be mitigated with feature importance methods [2]. The one used in this thesis is permutation feature importance. This a process by which the standard model prediction is compared to the same model but with one of the variables randomly shuffled. The loss in prediction from the standard model to the randomly shuffled model is

the importance of that feature and is given as a positive value [44]. Note that if the feature is irrelevant it will report near zero value or even a slight improvement on prediction strength due to random fluctuation. Features can have a negative predictive importance, in which case including them in the model reduces the model accuracy [77]. Some of the features used in the model in this thesis are categorical features these are changed into binary variables, so the day of the week column had values for Monday-Sunday, and this was changed into seven separate columns of if Monday value=1 if not value=0. This method is called one-hot encoding [81]. These are later re-combined after being used for prediction to allow for better comparison. A drawback of permutation feature importance is that it can be computationally expensive to compute.

The metric that will be predicted is delay against the timetable. The variables given to the regression model to predict this delay are direction of travel, TIPLOC of arrival, time of arrival (given as seconds past midnight), day of the week and type of train (see breakdown of headcode in chapter two).

The data set spans five weeks from September 1st, 2019, to the 6th of October 2019. This data was selected to be within one six-month timetable. This allows all the days to be compared. It also features no bank holidays; these prevent comparison due to different schedules and staffing on such days. They also do not fall within the summer holidays, or the Christmas break so are representative of typical operating conditions. Bank holidays operate similar to Sundays with less trains and less delays. Different six-month timetables would operate in a similar manner to the one being examined in this thesis. The aim of this work was to look at a commuter line that was of interest to our industrial partner that had sub-threshold delay issues. These at the time (before the Covid-19 pandemic) considered the priority of Network Rail.

One week is held back for validation purposes after fitting. Then the data is processed to prepare the variables for the regression. Outliers that are more than ten minutes late or early are removed, this represents less than two percent of the data set and drastically improves prediction quality. One hot encoding is used to transform the categorical variable into binary representations in different columns allowing fitting to occur. This handling of categorical variables is why a random forest approach was chosen.

The prediction errors for the whole line are 51.14 seconds of mean error and 36.78 seconds of median error. The mean error being higher shows that despite the removal of the worst outliers there remains a clear bias towards higher amounts of delay showing more influence.

The errors found when predicting the validation data set were slightly higher

with 64.3 seconds of mean error and 39.4 seconds of median error. This shows that there was a slight overfitting problem but not that significant for the median error. It could however just be down to randomness between the datasets or a more disrupted week of data.

The next action to take is to measure which of the features that we have given to the model are utilised the most. This is done by calculating the feature importance.

Feature	Importance value
Time of day	0.464
TIPLOC	0.447
Direction	0.227
Day of week	0.179
Service Type	0.170

Table 3.1: Permutation importance values for the whole line.

Table 3.1 shows the values for the whole line with the time of day being the most influential measure. This is not surprising because section 3.4 shows that peak times can drastically increase delay. TIPLOC of arrival is the next most significant measure. This highlights the role that the infrastructure plays in the delay on the line. This can also be seen in figures 3.7 to 3.11. Direction, another infrastructure measure, is next in significance. This is interesting since our understanding from the observations we made is that direction needs to be combined with time of day variations to have an effect. This is because of the morning rush hour effect which should increase delay towards London Fenchurch Street. The opposite effect should also occur in the evening. It is possible that this measure would have more significance if varied with time of day as well. Day of week is moderately important due to differences between weekday and weekend schedules. The latter are less congested so suffer less from chronic congestion. The last feature is service type which is the difference between types of services and even the engines used to pull them. This variable is moderately useful in determining the delay as well.

We now shift focus to individual TIPLOCs on the line. The output of the regression method shows that predictions at Barking had an average mean error of 63.04 seconds and an average median error of 52.97. The main objective of this regression method was to measure the feature importances. This would allow us to

see the impact of the network features on the delay of the trains. The importance of the features on different TIPLOCs on the line are shown in table 3.2.

TIPLOC code	Time of day	Day of week	Direction	Service type	Sample Size
FENCHRS	0.145	-0.043	0.179	-0.033	9075
LIMHSE	0.454	0.220	0.215	-0.005	17332
WHAMHL	0.238	0.160	0.160	-0.014	18335
BARKING	0.297	0.134	0.082	0.065	18420
UPMNSTR	0.376	0.217	0.120	0.062	13662
WHORNDN	0.325	0.213	0.195	0.019	4974
LAINDON	0.332	0.172	0.155	0.04	8961
BASILDN	0.299	0.137	0.282	0.139	8204
PITSEA	0.459	0.101	0.121	0.292	11213
BENFLET	0.330	0.200	0.414	0.058	11321
LHONSEA	0.348	0.215	0.136	0.105	11389
WCLIFF	0.336	0.012	0.329	0.012	10798
STHCENT	0.340	0.181	0.282	0.04	9775
STHNDE	0.340	0.181	0.246	0.076	8283
THOPBAY	0.279	0.149	0.529	0.066	8286
SHBRYNS	0.271	-0.019	0.284	-0.025	4096

Table 3.2: Feature importances for individual TIPLOCs on the Fenchurch St to Shoeburyness line.

This model was also used to predict delays arrivals at other TIPLOCs on the line as shown in table 3.2. The table shows us that by far the most useful predictor of delay is the time of day. The direction of travel is the next most significant indicator followed closely by the day of the week on which this arrival was recorded. The service type of the train was shown to be mostly useless.

The sample size column shows much of the topology of the line itself. London Fenchurch St and Shoeburyness are the termini on the line so don't have trains going through the station which halves the observations. Many trains depart the

line at Barking as well as some leaving at Upminster to travel along the branch line. Some of these return at Pitsea followed by some trains terminating early at Southend Central on the two stopping platforms. The lower observations at West Horndon are caused by faulty berths at that location which do not detect every train that passes through. This method shows which TIPLOCs are more unpredictable and whether various times of day can affect delay statistics. A box plot of these importances at Barking is displayed in figure 3.13.

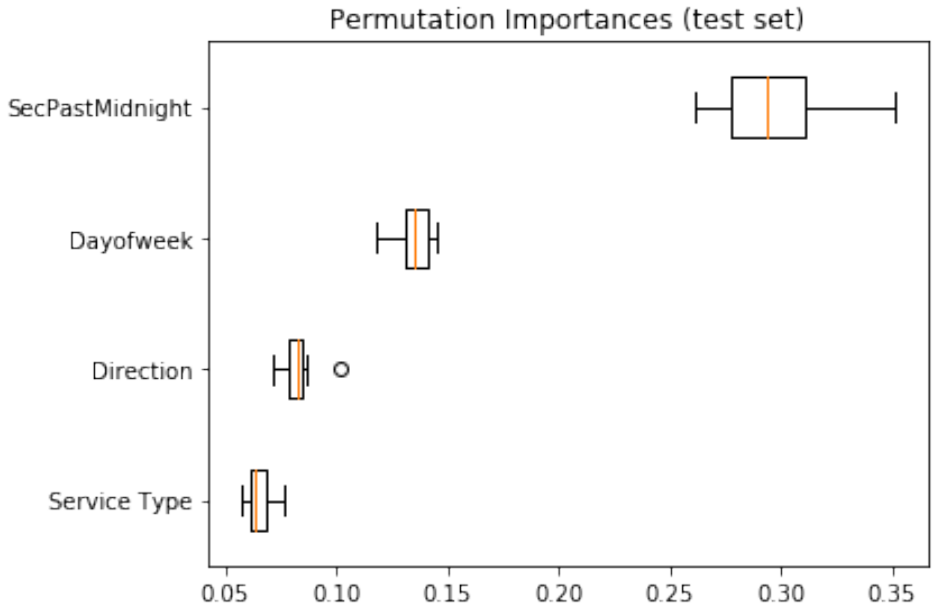

Figure 3.13: Boxplot of feature importances at Barking.

3.6 Conclusion

This chapter and the corresponding appendix have shown the amount of effort required to clean this data and the processing needed to allow it to be fit for analysis. This cleaning is unfortunately mandatory and is a barrier to many people trying to get into Network Rail data analysis. If this were simpler it would allow for easier access to this data for the general public, as was intended by Network Rail, but also for the purposes of academia. Our work in this area was only achieved with cooperation from industry insiders and supplementary data such as the STANOX to berth mapping.

The SRTs are introduced in this chapter for extensive later use. They allow for more interpretable results such as finer-grain location of trains within a berth. We have also looked at a capacity model for the mainline calculated with the SRTs as well as explanations for the values given.

Bottleneck analysis for the line can help to determine structural issues on the line and our analysis found areas that can be improved. Further work in this area could be seeing what the new bottlenecks are after the existing ones are fixed.

The random forest approach showed us the importances of the different features to see the influence of the infrastructure on the line. If further work were to be done in this area, we would add some more features and attempt to work on a larger set of data. It would also be interesting to see how the model performs on a new timetable. Additional tweaks could be made to improve this model in the future and it will be adapted for use in chapter four. We improvised to get time of day as a metric into this model. The seconds past midnight measure performed better than expected, as measured by the feature importance. This could be of use to future researchers in this area. Algorithm performance was not a focus for this thesis, but adaptability was. Future analysis could look at the important infrastructure characteristics on other national networks.

Chapter 4

Analysis of Services on a Line

4.1 Introduction to services in the line

In this chapter we will be examining the services that operate on the London Fenchurch St to Shoeburyness line. It will be examining the interactions between these trains as well as discussing the operational side of the railway line. This is the partner chapter of chapter three but moves the focus from the infrastructure to the trains that travel on it.

Traffic management must balance capacity against delay, since the more trains that are added to a network the less reliable the services are [62]. The utilised capacity of the network is proportional to train-on-train delay. Delay here is measured in seconds behind schedule. The work here will examine the delay caused by this high-capacity line and suggest improvements to the operations. The solutions proposed will avoid suggesting a reduction in services. Instead, we will propose predictions to inform operators earlier of incoming delays. This will allow for better handling of the disruptive trains.

4.2 Travel times between TIPLOCs

Timetables on the Great Britain railway are not always reliable and are planned more than a year in advance of use. Most of the timetable compilation is done manually so lacks efficiency [49]. We would recommend reading Kessell's work here if you would like a better understanding of the timetable construction process. Kessell also describes challenges that need to be overcome to improve the current situation, the first of which is automated production of timetables (part of which would be checking viability). Previous works in the area of timetable automation such as Sels

et al [82] discuss the limited application of academic timetable automation. Working with and assisting existing timetable construction could lead to more application in practice. Here we suggest a method of timetable validation. Travel times between TIPLOCs can be used to test the plausibility of timetables because they provide averages for the real traversal time. This can then be used to inform allowances that are added to the timetable. Allowances are additional minutes of leeway that can be added to the timetable to increase reliability, they can be found in the schedule data (Chapter 3.2).

The approach used by us to find these travel times is to use the combined timetable and berth data to determine accurate arrival times for different TIPLOCs. We then track services to monitor which TIPLOCs they stop at and when they do so. This is applied to thirty-six days of data comprising of over 228,000 arrivals at TIPLOCs.

These arrivals are then matched by TIPLOC pair then the time subtracted and averaged over all of the journeys. Offsets are included here because we are comparing the berths to the timetable. Offsets, as mentioned in appendix A.4 are a method of translating berth arrivals to platform arrivals. We then calculate the median and mean travel time between TIPLOCs on the line.

TIPLOC pair	Median Travel Time	Mean Travel Time
FENCHRS-LIMHSE	00:04:07	00:03:57
LIMHSE-WHAMHL	00:04:23	00:04:21
WHAMHL-BARKING	00:04:16	00:04:16
BARKING-UPMNSTR	00:07:18	00:08:18
UPMNSTR-WHORNDN	00:04:10	00:04:06
WHORNDN-LAINDON	00:04:12	00:04:09
LAINDON-BASILDN	00:02:01	00:01:53
BASILDN-PITSEA	00:03:10	00:03:12
PITSEA-BENFLET	00:03:05	00:03:05
BENFLET-LHONSEA	00:03:47	00:03:45
LHONSEA-WCLIFF	00:04:26	00:04:22
WCLIFF-STHCENT	00:01:56	00:02:06
STHCENT-STHNDE	00:01:38	00:01:31
STHNDE-THOPBAY	00:02:04	00:01:58
THOPBAY-SHBRYNS	00:03:46	00:03:58

Table 4.1: Travel times between TIPLOCs of the Fenchurch St to Shoeburyness line Eastbound.

We now compare these travel times to an example of a train scheduled to take this journey to verify the timetable is plausible to be done. An example service, with the headcode "2B76" is shown below:

Message type	TIPLOC	Arrival time	Departure Time
LO	FENCHRS	0005	0005
LI	LIMHSE	0009	0009H
LI	WHAMHL	0014	0014H
LI	BARKING	0019	0020
LI	UPMNSTR	0027H	0028H
LI	WHORNDN	0033	0033H
LI	LAINDON	0037H	0038
LI	BASILDN	0040H	0041H
LI	PITSEA	0044H	0045
LI	BENFLET	0048	0048H
LI	LHONSEA	0052H	0053
LI	WCLIFF	0057	0057H
LI	STHCENT	0059H	0100H
LI	STHNDE	0102	0102H
LI	THOPBAY	0104H	0105
LT	SHBRYNS	0109	0109

Table 4.2: Schedule for headcode "2B76".

We can refer back to section 3.2 to see the explanation of the various fields in this table but for our work here we need only recall that a "H" after a time means that time plus thirty seconds.

Now we will test the plausibility of this timetable since allowances are in increments of one minute, that is how we will measure how much the schedule deviates from the travel times. The first step in the timetable takes four minutes and thirty seconds. This is plausible when compared to the median travel time of the same section of track.

In the same way each of the differences are calculated and the largest difference is the measurement between LAINDON and BASILDN which takes twenty-nine seconds less time than the timetable has planned. The largest difference in the other direction is twenty-six seconds which is between LHONSEA-WCLIFF. The conclu-

sion here is that according to median travel times calculated, there are no allowances that are required for this timetable. The only one allowance needed when compared to the mean travel time would be between Barking and Upminster but the mean measurement could be skewed here by outlier values.

An interesting observation here is the time set aside for stopping at the station seems to be thirty seconds at smaller stations and sixty at larger ones. This is likely because of historic stopping times at these locations. An obvious outlier here would be Basildon where the longer stopping time is applied. This is likely due to a timetable solution such as having a thirty second allowance disguised as extra stopping time. Or is related to the junction at Pitsea.

This method can form part of the solution to the challenges set out by Kessell [49]. With more work at this we can continue to strive for an improved timetabling method. This will reduce the amount of effort that needs to be made by hand and also could give a superior end-product.

4.3 Introduction to train graphs

This next section will be looking at how services can be displayed and compared to one another. The method we decided on was a train graph and an example is shown in the 2008 paper Generating Train Plans with Problem Space Search [74] where it is used to visualise a train plan.

This choice allows timetables to be encoded as lines on a distance-time graph. Since these graphs display actual services, they can only show one day of data at a time. Multiple days would add unnecessary clutter and remove all interpretability.

Train graphs vary the axis in the literature, some plot locations on the x-axis [56] and others on the y-axis [65]. In this thesis the y-axis will display the TIPLOCs in order of where they are on the line, they are currently placed at equal distancing apart. The x-axis will be time. The data is sorted by headcode to allow each one to be plotted separately. Direction is also used to filter the trains, both directions can be plotted at the same time, but they have little interaction because this line is bi-directional and there are almost no services that travel in both directions. There is an exception here, a single service travels in both directions on the line. This is a class three train. This means it is a priority empty carriage, parcel train or seasonal train (weather dependent). The most likely option is a parcel train here because it travels back and forth across the line.

TIPLOC	2B50	2D00	2D02
FENCHRS	2019-10-01 05:00:00	NaT	2019-10-01 05:07:00
LIMHSE	2019-10-01 05:04:00	NaT	2019-10-01 05:11:00
WHAMHL	2019-10-01 05:09:00	NaT	2019-10-01 05:16:00
BARKING	2019-10-01 05:14:00	2019-10-01 04:53:00	2019-10-01 05:22:00
UPMNSTR	2019-10-01 05:22:00	2019-10-01 05:01:00	2019-10-01 05:30:00
WHORNDN	2019-10-01 05:27:00	NaT	NaT
LAINDON	2019-10-01 05:32:00	NaT	NaT
BASILDN	2019-10-01 05:35:00	NaT	NaT
PITSEA	2019-10-01 05:39:00	2019-10-01 05:35:00	2019-10-01 06:04:00
BENFLET	2019-10-01 05:42:00	2019-10-01 05:39:00	2019-10-01 06:08:00
LHONSEA	2019-10-01 05:47:00	2019-10-01 05:43:00	2019-10-01 06:13:00
WCLIFF	2019-10-01 05:51:00	2019-10-01 05:48:00	2019-10-01 06:17:00
STHCENT	2019-10-01 05:55:00	2019-10-01 05:51:00	2019-10-01 06:20:00
STHNDE	2019-10-01 05:57:00	NaT	NaT
THOPBAY	2019-10-01 05:59:00	NaT	NaT
SHBRYNS	2019-10-01 06:03:00	NaT	NaT

Table 4.3: Schedule for first three trains on October the 1st 2019 on the Fenchurch St to Shoeburyness line.

The three trains seen in table 4.3 show three different journeys that can be taken along the line. We can see that this is the schedule due to the round numbers at the TIPLOCs. The first service, headcode "2B50", is the most standard in that it starts at London Fenchurch Street and progresses smoothly through all the TIPLOCs and terminates at Shoeburyness. The next train starts at Barking then progresses to Upminster where it leaves the mainline for the Tilbury loop and does not return until Pitsea. It then finishes at Southend Central. We can observe that the Tilbury loop takes longer to traverse than the mainline since the difference between train one and two at Barking is twenty-one minutes and this has shrunk to four minutes by Pitsea. The last service shown here is the same as the second one but originates at London Fenchurch Street before travelling to Barking then

following the aforementioned route.

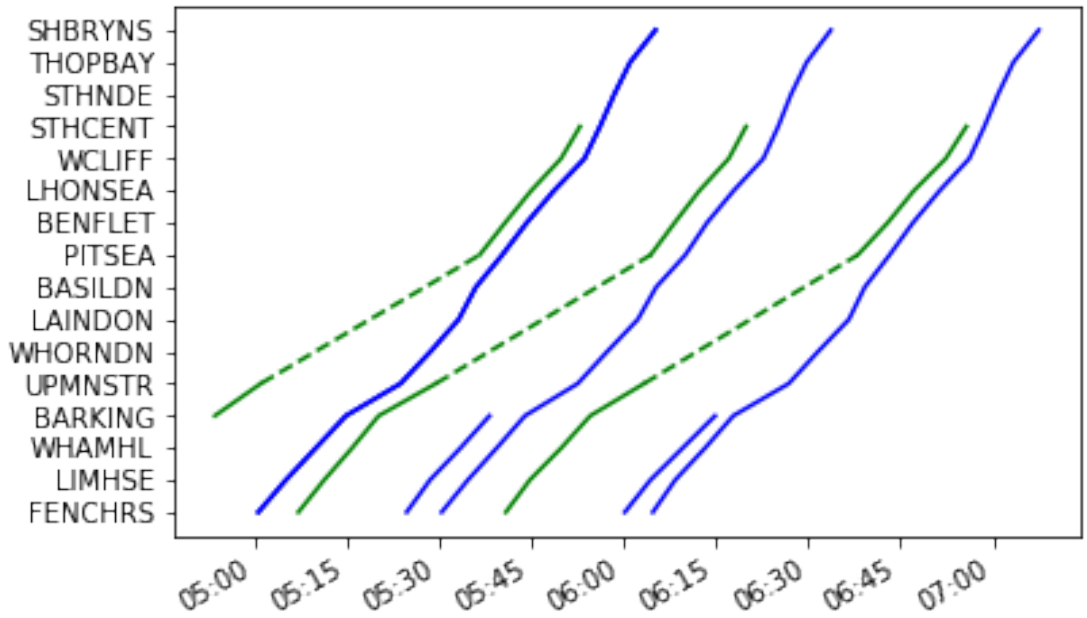

Figure 4.1: Train graph eastbound for the first eight trains of 01/10/2019. on the Fenchurch St to Shoeburyness line.

Figure 4.1 shows a train graph of the first eight trains seen on the 1st of October, using 4am as the start of the day. The line plotted in green are ones that miss at least one TIPLOC out before rejoining the line. The discontinuous periods whilst the service is on the Tilbury loop are plotted as dashed lines linking up the two parts of the service on the mainline. The lines cannot intersect because overtaking is not generally done on this line and is certainly not planned to be done in the schedule. However, the trains can overtake if one of them is on the Tilbury loop. This means that two lines can cross but only if one of them is dashed.

Interestingly, the lines have a minimum distance apart best shown at 6:10 at Barking where two trains approach this distance.

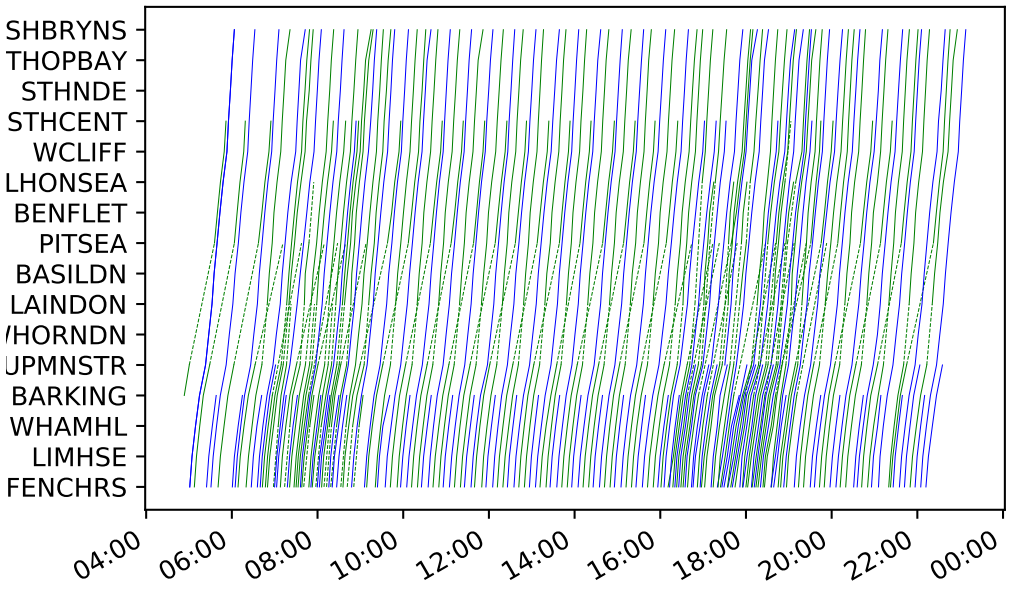

Figure 4.2: Tuesday 1st October 2019 eastbound scheduled trains. Plotted in blue for mainline trains, green for train the enter the branch line.

Figure 4.2 shows the full eastbound schedule plotted for the same day. By the density of the lines together we can see the peak times of 8am for the morning commute and the evening peak between 5pm-7pm.

4.4 Train interactions (concertinas)

This section will explore and explain the minimum distances apart discussed in the last section. The train graphs seen in figures 4.1 and 4.2 are plotted separately with no notion of relationship between trains. We can see from the graphs that these trains are not independent.

Therefore, we developed a model that can measure these interactions and report locations of trains on the line. It works on a historical data set and cannot yet be applied in real-time but could be easily adapted for this purpose.

The first stage of monitoring where the trains are in the network is to have a data frame of the locations of the trains at any given time. We construct a list of berths from one end of the line to the other in both directions. This is done by

tracking trains that traverse every TIPLOC on the route.

Then TIPLOCs are added to their respective locations and so is a column for the occupant of the berth, the observed time of that train at that berth, and the last time a train was detected at that berth.

Berth	TIPLOC	Occupant	Observed Time	Last Time
C354	NaN	NaN	NaN	2019-09-30 05:25:40
A354	SHBRYNS	NaN	NaN	2019-09-30 05:28:42
0352	NaN	NaN	NaN	2019-09-30 05:29:56
0350	NaN	NaN	NaN	2019-09-30 05:30:27
0348	THOPBAY	2B57	2019-09-30 05:30:56	2019-09-30 05:15:58

Table 4.4: Sample of the west berth list dataframe.

Table 4.4 shows an example of the state of the first five berths westbound, the dataframe is updated as more berth movements are observed throughout the day. This snapshot was taken at half past five in the morning of the 30th of September. It shows a train present at Thorpe Bay and a comparison against when the last train was in that berth. You can see that as a train departs a berth it leaves behind the time it arrived at that berth in the "Last Time" column. This allows us to compare headway as a measure of time. There is also a table that updates the current situation for each train. It is blank initially and trains are added to it as they are observed on the line. The berth list dataframe stores the current properties of that train.

Headcode	Direction	Last time	Status	Berth	Last TIPLOC	Prior Train	Post train	Re-occupancy time	Headway

Figure 4.3: Headcode table shown in its initial state.

As can be seen in figure 4.3, which depicts a table, there are more attributes tracked for a train than a berth. The unique identifier of choice here is the headcode because of its presence in both the berth and timetable datasets. Of note here are the Prior train and Post train columns. They show if a train is within four berths of the original service. Prior train if the other headcode is in front and post train if the

train is behind. Four berths is the chosen separation because three berths apart is the threshold at which the further back train needs to potentially slow down due to the train at the front. However, trains can vary between two and four berths apart whilst close to one another so four berth observation allows for the situation to be continually monitored. This could sometimes register trains that never approach within three berths and that is a drawback of this monitoring method. The rear train in this relationship must slow down to compensate for the lack of headway. This interaction can be measured as knock-on delay. Trains can be as few as two berths apart, but this is subject to harsher speed restrictions and requires the driver to go through a caution signal. Due to the way that the data we are using reports, we receive messages after each berth transition. This means that based on which report occurs first, trains three berths apart can vary between two and four berths apart. Since berths are not of uniform length this can happen frequently. This relationship between trains is what will henceforth be described as a "concertina". Named after the musical instrument that can compress and expand yet still remains attached. Trains in this relationship are referred to be "in concertina". Trains are considered to be no longer "in concertina" if one of the following situations occurs:

1. One of the trains terminates at destination
2. The leading train accelerates out of the concertina
3. The following train falls too far behind
4. One of the trains leaves the mainline

The last of the prepared tables is shown below:

Headcode1	Headcode 2	Start time	End time

Table 4.5: Initial train interactions (concertina) dataframe.

It shows the list of concertinas on the line and when the trains were in this relationship. The above tables are updated throughout the day of data depending on how far into the day the model has progressed. The model advances through the day in time order treating each berth movement separately. They are first checked to see if they are movements covered by the mainline. If this is confirmed all other mentions of that headcode are removed from the west berth list data frame. The train is then added at the new location. The next step is to check the headcode

table and update each column. This stage is where concertinas are checked and added before or after if necessary. The distance checked is four berths either side to match the maximum distance that concertinas can vary by. The last steps are checks for removal of the train either by arriving at its destination or if it leaves the mainline. This process is then repeated in the eastbound direction. And for all movements for the day of data being processed.

Each of the separate concertina observations are processed afterwards into the various time ranges of the relationships. This is done by grouping with headcodes. An assumption made here is that two headcodes can only enter one concertina over the course of the journey with each headcode. So, if a headcode "1A00" were to enter a concertina with headcode "1A01" at 9:00am until 9:10am and a separate concertina between 9:20am and 9:30am, this model would assume that the headcodes were in concertina from 9:00am to 9:30am.

Concertinas are pairwise interactions but can form in chains with services being in concertina with multiple trains simultaneously. This can happen due to a train having concertinas with both a train in front and a train behind.

For example, on the 1st of October 2019 at 8:47am there was a concertina of length six present in the data. This shows that in the real world there are significant numbers of trains that closely follow one another through the network. Concertina interactions can be thought of as an interval graph where the nodes are headcodes and the edges are connections between services that are within four berths.

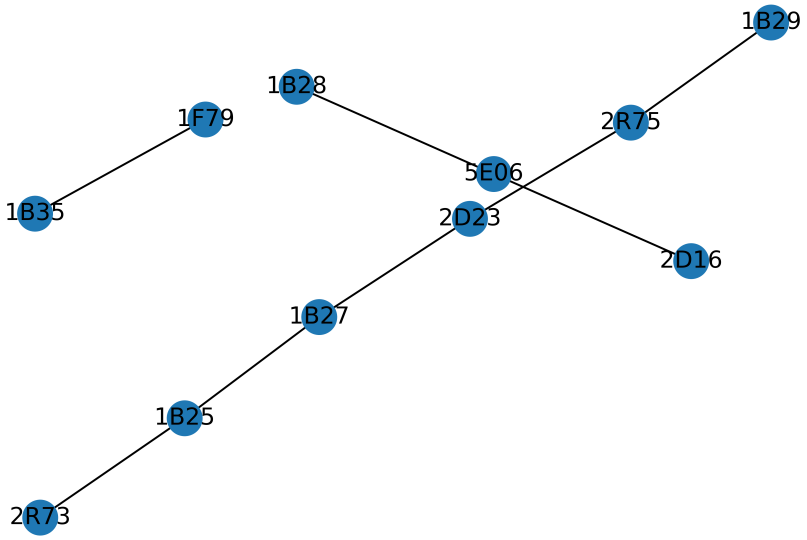

Figure 4.4: Train interactions (concertina chains).

Figure 4.4 shows the concertinas present in the data at a particular minute mid-morning on a Tuesday. This is during rush hour so there are more trains on the line than at other times of the day. The longer chains do not come into existence instantly however, but slowly grow as more trains join the end of the concertina. In the same way they shorten by trains leaving the front of the concertina mostly by terminating at their respective destinations. This is how the concertina of length six shown in the figure broke down.

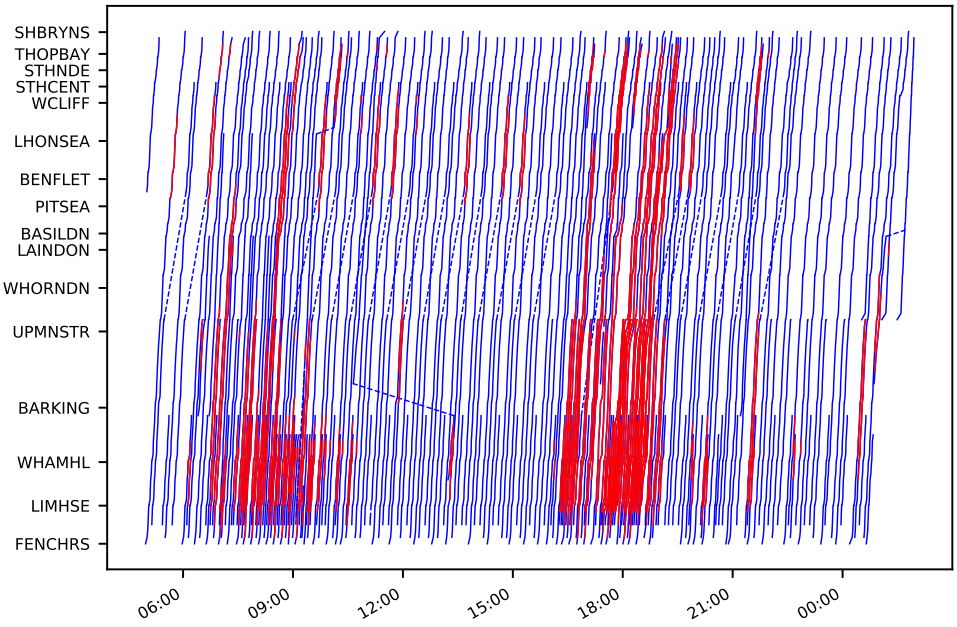

Figure 4.5: Eastbound train graph for Tuesday 01/10/2019.

Figure 4.5 shows the eastbound train graph with the same day as figure 4.2. The difference here is that these are the real movements instead of the scheduled movements. The trains are plotted in blue unless they are in a concertina at that time, in which case, they are red. The dashed blue line seen travelling westbound between 10am and 1pm is a parcel train that is travelling in both directions on the line.

This graph shows that, at least on this day, concertinas tend to occur nearer to Fenchurch St where the trains are more frequent. They are also more likely to happen during the rush-hour periods, with more severe congestion in the evening

peak.

4.5 Knock-on delay

Knock-on delay is the idea that one delayed train service can delay other services by obstructing them from proceeding on the track. The method we will be examining them by will be through the concertinas seen in the last chapter. Concertinas are unplanned interactions that can transfer delays from one train to another.

Delay against the timetable can only be monitored at TIPLOCs for each service. However, using this approach would not allow us to get an idea of how delay was evolving between stations. Section 4.1 shows us that updates on delay could be as infrequent as seven minutes apart for the longest section between TIPLOCs.

We proposed a method that updates delay at each berth by comparing the time taken to traverse each berth against the sectional running times calculated in chapter three. So, if a train takes ten seconds longer than the expected median time to travel between two berths, that time is added to the delay measurement. This is done for each intermediate berth up until the next TIPLOC where our estimate is overwritten by the next measurement against the schedule. This method is not perfect because most delays experienced by trains can come from dwell times at stations.

This estimated delay allows us to now have a delay measurement for every berth movement.

Again, the train graph will be used to visualise the problem at hand.

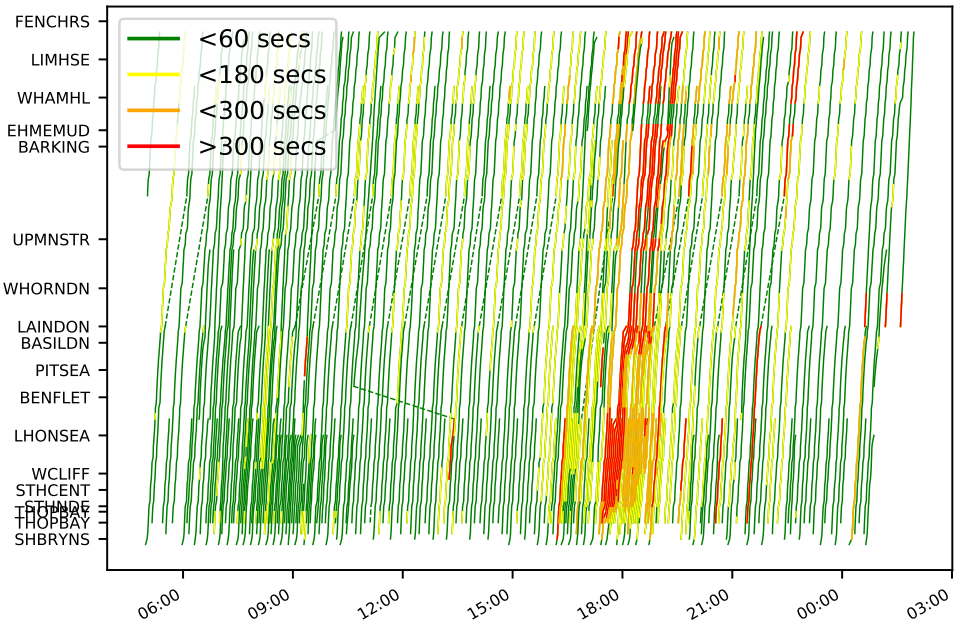

Figure 4.6: Eastbound delay train graph for Tuesday 01/10/2019. Trains drawn in green are less than 60 seconds late. Yellow denotes between 60 and 180 seconds late. Orange for delays between 180 and 300 seconds late and red for delays in excess of 300 seconds.

Figure 4.6 shows the same graph as figure 4.5 but the colour scheme this time is based on how delayed the service is. Green signifies on time, defined here as less than one minute late. Yellow shows trains that are between one and three minutes late. Orange trains show trains that have delays above the threshold for investigation by the Delay Attribution Board, meaning greater than three minutes and in this case less than five. Red shows the trains that are in excess of five minutes behind schedule. There is clearly a substantial delay on the line just before six in the evening. A close-up of this incident is shown below:

Figure 4.7: Highlighted view of incident in train graph delay format. Trains drawn in green are less than 60 seconds late. Yellow denotes between 60 and 180 seconds late. Orange for delays between 180 and 300 seconds late and red for delays in excess of 300 seconds.

The incident begins when a train that is departing late enters a concertina with the next service forcing it to also be delayed. The original inciting train actually recovers as it travels down the line and gets back on schedule by Shoeburyness. This does not occur to the other trains it leaves behind. These services are delayed by each proceeding train leading to a cascade of delay impacting all the trains that

depart over the next half-hour. There are a few cancellations that occur at Barking, which improves the situation, but a further incident at Upminster with a leading train experiencing a larger than normal dwell time causes the network to again be delayed. This delay did not clear further up the line and many of the services over the next hour arrived more than five minutes late at Shoeburyness. The full effect can only be seen if we observe the Westbound direction.

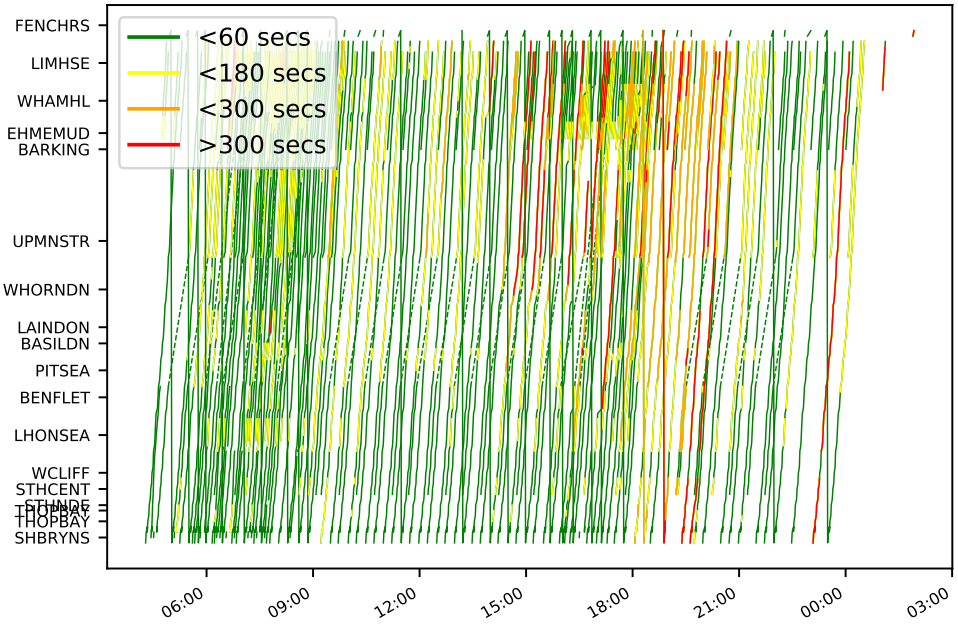

Figure 4.8: Tuesday 01/10/2019 westbound train graph showing delay. Trains drawn in green are less than 60 seconds late. Yellow denotes between 60 and 180 seconds late. Orange for delays between 180 and 300 seconds late and red for delays in excess of 300 seconds.)

Figure 4.8 shows us the causes and effects of the incident described above. The causes are seen by the trains in red approaching Fenchurch St at the top of the graph. These are the same trains that after turning around at the end station depart late. The effects are seen at Shoeburyness at approximately 19:00 with trains departing in excess of five minutes late.

This graph shows clearly that the warning signs of the incident were seen earlier in the day than were obvious from the eastbound graph. So east and west

on this line are linked by turn-around times at terminus stations. These terminus stations act as delay dampeners by absorbing some of the delay. However, if these were to be increased the delays in the other direction could be avoided.

We could examine from these graphs why the westbound trains were delayed (seemingly caused by trains joining from the Tilbury loop) but this job is for the Delay Attribution Board who decide the root-causes of delays for the purpose of allocating compensation payments.

More examination of these turn-around times in more detail. Unfortunately, there is no guaranteed way of tracking trains that come in and then come out again. This is because upon completion of their route, the train changes headcode and trainUID. Despite this, we can count trains in and out of specific platform berths. This is not perfect since platforms can store multiple trains, but it gives a high-level of certainty.

The data we chose was the 1st of October since it is displayed above in figure 4.6. We will examine the Shoeburyness turn-around due to it having less platforms. From observing the maps found on Open Train Times [41] we can identify which berths correspond to the platforms. As there are three platforms there are also three inbound berths and the same number outbound. The data is sorted to obtain the berth movements associated with these berths and then the data is paired with respective inbound and outbound berths. This data is sorted chronologically and adjacent in and out movements are paired. This leads to the following table:

Scheduled Arrival	Scheduled Departure	In Delay	Out Delay	Scheduled Stopover
2019-10-01 04:05:00	2019-10-01 04:59:00	00:02:32	00:00:17	00:54:00
2019-10-01 05:21:00	2019-10-01 05:28:00	-00:02:27	00:00:14	00:57:00
2019-10-01 05:35:00	2019-10-01 05:44:00	-00:03:06	00:00:26	00:09:00

Table 4.6: Turn-around pairs Dataframe

Values with extreme In/Out delays are removed which allows us to then plot the values of inbound delay against outbound delay. This is shown in figure 4.9.

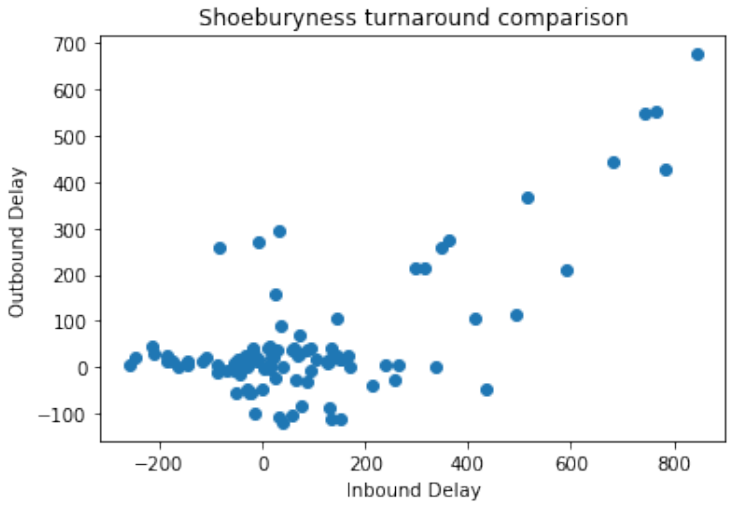

Figure 4.9: Tuesday 01/10/2019 Shoeburyness inbound delay against outbound delay.

Fitting a simple linear regression to this gives us a slope of 0.443 and an intercept of 13.93. If fitted for an inbound delay of less than 300 seconds we now get a slope of -0.01 with a similar intercept of 14.75. This shows that with less than five minutes of inbound delay there is no discernible relationship between inbound and outbound delay. However, if we do the same fit for values in excess of 300 seconds of delay the slope becomes 0.94 with an intercept of -216. So, there are clearly two patterns of behaviour here. A formulation can be made to model this behaviour where

$$\begin{aligned} \mathrm{O} &= \text{Outbound delay}, \\ \alpha &= \text{Scheduled Stopover}, \\ \mathrm{I} &= \text{Inbound Delay}, \\ \gamma &= \text{Minimum Turnaround}. \end{aligned} \tag{4.1}$$

Then the relationship can be described as

$$\mathrm{O} = \begin{cases} \alpha \geq \mathrm{I} + \gamma & \mathrm{O} = 0, \\ \alpha < \mathrm{I} + \gamma & \mathrm{O} = \mathrm{I} + \gamma - \alpha. \end{cases} \tag{4.2}$$

There is also additional noise to take account of in real-world applications,

so this formula is not perfect. Based on the above data set we can assume that the minimum stopover is approximately five minutes.

The conclusions here are the way that these train graphs can be used to examine historical incidents and learn how to prevent such disruptive incidents from occurring in the future. For example, utilising the five-minute stopover to better schedule turn-around at Shoeburyness. Allowances like the timetable has could be added to these services at their terminuses. Correct implementation of this would mitigate or prevent delay travelling in the other direction on the mainline. This day of data was particularly disruptive. With a cumulative positive delay at TIPLOCs of 10 days and 12 hours for passenger trains. This represents 32.0 seconds of average delay per train. Below is a more typical day on the Fenchurch St to Shoeburyness line with only 23.6 seconds of average delay per passenger train:

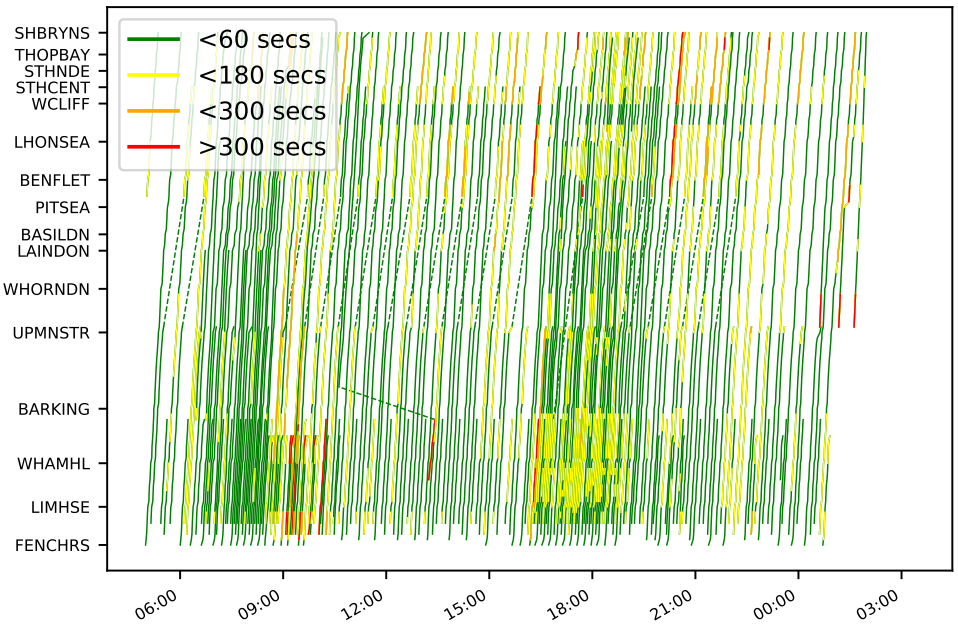

Figure 4.10: Thursday 01/0/2019 eastbound train graph showing delay. Trains drawn in green are less than 60 seconds late. Yellow denotes between 60 and 180 seconds late. Orange for delays between 180 and 300 seconds late and red for delays in excess of 300 seconds.

Weekends, especially Sundays, are far less congested and proceed as sched-

uled most of the time. This Sunday had 11.8 seconds of average delay per passenger train.

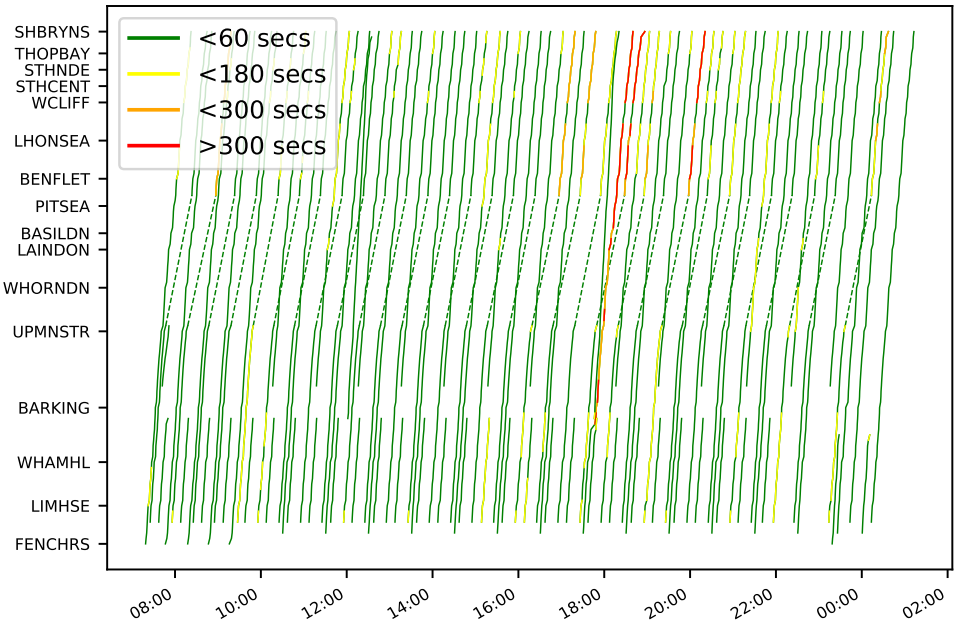

Figure 4.11: Sunday 29/09/2019 eastbound train graph showing delay. Trains drawn in green are less than 60 seconds late. Yellow denotes between 60 and 180 seconds late. Orange for delays between 180 and 300 seconds late and red for delays in excess of 300 seconds.)

Turn-around times were also examined at the two turn-around platforms at Southend Central. The study here found that this station handled delays in a similar manner but there were less delayed trains entering and near zero lateness when departing. This could be because the delay hasn't yet reached the levels required to export delay in the other direction. Another explanation is that these platforms have less utilisation and this helps mitigate delay problems. A study was also done looking at the platforms at London Fenchurch St station. Due to the complex topology of the station, matching inbound and outbound trains was not possible. More work could be done here to look at terminuses in other areas of the network.

4.6 Prediction of arrival times at stations and travel times

This section will examine the prediction of arrival times at stations. Various methods will be explained and then used to predict arrival times at various TIPLOCs on the line. The baseline that will be predicted against is the timetable itself. It is publicly available and lists arrival times at each station. If a prediction method cannot beat the timetable, then it is not good enough to be implemented.

The methods will be compared as absolute deviations from reality in seconds. Therefore, predicting arrival times earlier or later than actuality are penalised equally. Predictions are done on a basis of time horizons. For example, a ten-minute time horizon would represent what arrival time would have been predicted, if that prediction were to have been done ten minutes before arrival. The timetable is not time dependent in its "predictions" therefore, it will not vary based on the time horizon. This does not apply to the other prediction methods proposed here.

The first such prediction method utilises the sectional running times (SRTs) discussed in the last chapter. This prediction requires pre-calculated SRTs as well as berth movement data from the same weekday of the preceding week. The first step we look at is to sort the data for trains that arrive at the target TIPLOC. Then the observation of when the train arrives at the target TIPLOC is found, the time horizon is taken into account to find the berth location that we will be predicting from. The next step is to locate that berth in the prior day's data set as well as the set of berths between the prediction berth as well as the berth that the TIPLOC corresponds to. The SRTs for the intermediate berths are then added to the arrival time at the prediction berth, then that sums up to our prediction for this train arriving at that TIPLOC with the chosen time horizon. The prediction result is compared to the actual arrival time and the absolute deviation from reality is recorded for later analysis. This is then repeated for all the services on that day which arrive at that TIPLOC. This method is time dependent because the prediction takes into account the time horizon and varies based on how many berths away the prediction is done.

The second such method is similar to the last method but with an extra step to further refine the use of the SRTs. The SRTs are first sorted by classification of the type of train. So express trains are split from all-stoppers and freight is treated differently as well. There is an additional step during prediction to select the correct SRT for the train headcode.

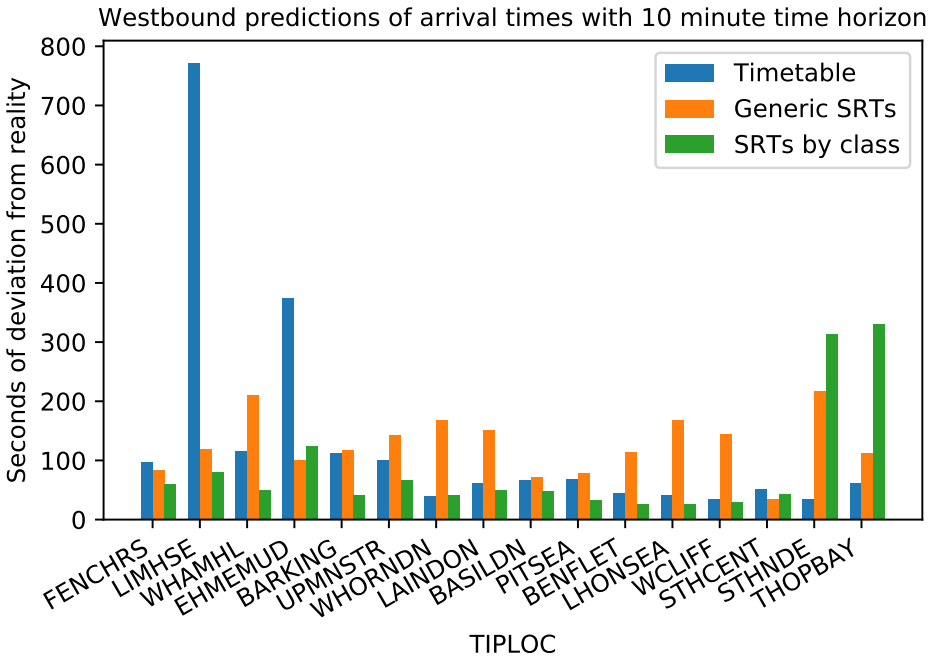

Figure 4.12: Comparison of prediction methods at each TIPLOC on the Fenchurch St to Shoeburyness line.

Figure 4.12 shows a comparison of the prediction methods introduced so far. The values are all means so the y-axis shows mean absolute deviation from reality. The first point of interest is the outlier values. The timetable predictions at Limehouse and East Ham Depot are particularly bad, this may be caused by poorly calibrated offset values that translate between the schedule and the berths. The SRT methods perform poorly at Southend East and Thorpe Bay because they are the first two TIPLOCs that are in the westbound direction. This means that the services being predicted did not exist ten minutes ago, causing the methods to perform poorly.

The worst method shown in this figure is the Generic SRT prediction which is consistently the worst performer. The best method is the SRT by class method due to it beating the timetable prediction at almost every TIPLOC. The Generic SRT prediction is inconsistent which stops it from achieving accurate predictions at all the locations.

The next model formulated is a linear regression model. It takes as inputs all the previously mentioned models as well as some other features. It is changed here

to instead predict travel times between the berth the prediction is done at and the TIPLOC at which you are arriving at. The method changes here from time horizons because otherwise the correct answer would always be approximately equal to the time horizon specified. Instead, we will now predict between five and ten berths away from the TIPLOC. This is done using random numbers and are generated for each day in the data set. The same set of random numbers are used for all the methods to allow for direct comparison.

The other features are the service type of the train (headcode leading number), The day of the week the prediction is on, the time of day and how many trains are in front of this train in a concertina. For the last measure we check if the current train is in a concertina at the time of the prediction. If this is the case, then it is checked that the train is at the rear of the concertina. The train in front is checked to see if it is also in concertina with a different train in front and this continues until the front of the concertina train is reached. So, the integer in this column represents how many trains there are in front of the train being predicted.

We possess five weeks of consecutive days of data. This is split up into training, testing and validation. Three weeks are used for training and one week for each of the others. This was not without its issues there were some data problems that had to be solved. Since the SRT methods require data on which movements will be made during the prediction period the data taken from the previous week must be similar enough to allow for comparison. An example of this can be seen with the data from the 15th of September, the third Sunday of our data set. On this Sunday the trains run a shortened version of all the routes, this removes TIPLOC arrivals towards the two ends of the line. Therefore, this day is not only hard to predict but cannot be used for train tracking purposes. The solution we decided on was that the prediction on the 22nd of September would instead utilise data from the 8th of September.

The data was fitted using the Scikit learn package in Python for both of the regression methods [72].

Prediction method	Mean absolute error	Median absolute error	Standard deviation
Schedule	1874.27	61.50	191.49
Generic SRT	77.79	14.00	55.41
SRT by class	44.91	12.00	45.31
Linear regression	45.89	19.77	44.69
Random forest regression	21.41	10.67	39.76

Table 4.7: Prediction method comparison ordered by standard deviation.

Table 4.7 shows a comparison of the prediction methods. The baseline prediction of the schedule is the worst performer of the methods with an excessive mean absolute error of half an hour. This is because the schedule does not have any indication of current track situations and cannot account for trains departing earlier or later than timetabled. This is caused mostly by non-passenger trains that have less emphasis on keeping to the timetable. The median here is a better representation of the schedule's prediction capability with an absolute error of just over a minute. This is closer to the expectation we had prior to the results. Due to the previously stated mean error issues the standard deviation of the prediction errors is also large.

The two SRT methods perform closer to expectation with mean absolute errors of 77.79 seconds and 44.91 seconds for the Generic SRT and SRT by class methods respectively. The significant difference demonstrates the difference between train types. The median absolute errors for the two methods are far closer at 14 seconds for the Generic SRT method and 12 seconds for the SRT by class method. These predictions are therefore significantly better than the baseline schedule prediction. The standard deviation is also larger for the Generic SRT method as compared to the SRT by class method. This is because of the same reason as the mean error difference.

The next method to review is linear regression. It performs worse than the SRT methods when it comes to the median absolute error at 21 seconds but is comparable with the mean absolute error. With all eight of the features given as inputs the linear regression method performs worse than the SRT methods. By careful feature selection; building from just the SRT by class input and rejecting features that worsened the prediction we finished by using just five of the eight features. This was done using the testing set and after the features were selected it was then run on the validation set. The discarded features were time of day,

day of the week and number of berths away from predicted TIPLOC. However, the performance of the linear regression dropped significantly between the testing data set and the validation one. The median in testing was 16.24 seconds and the mean was 30.50 seconds. This suggests that the method is over-fitted. The standard deviation here is however low so it has more consistent predictions than the SRT methods.

The last method we looked at was the random forest regression model. This model was significantly better than all previously addressed models with a mean absolute error of 21.41 seconds which is half of the next nearest method. It also has a lower median absolute error than the next best method of the SRT by class. It also has the lowest standard deviation so is the best methods by all metrics compared here.

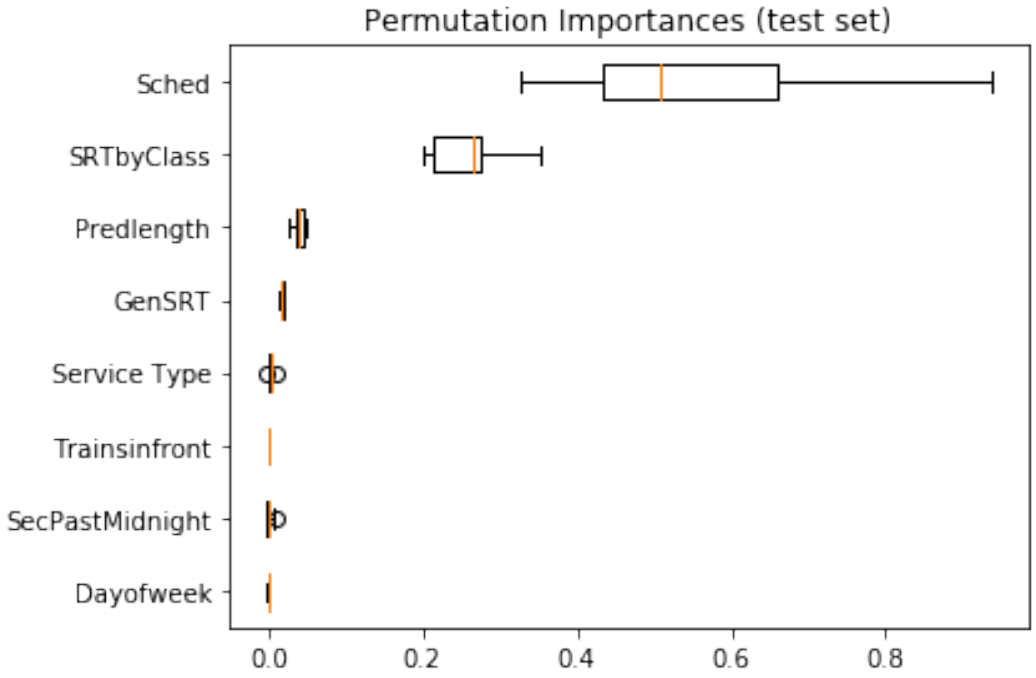

Figure 4.13: Comparison of feature importance for the random forest prediction method.

Figure 4.14 shows the importance of the features used by the random forest method. Surprisingly, given the values seen in table 4.7, the random forest method finds the greatest prediction value in the schedule and uses the SRT by class prediction as the next most valuable measure. Of the other features given, the most

utilised is the prediction length, here meaning how many berths away the prediction was done. The rest of the features had little use for the predictions. This means that the measure we devised to look at train interactions, number of trains in front in a concertina, was not utilised by the random forest. This is likely because 95.76 percent of all the trains in the testing data were not at the rear of a concertina so had a zero in that column. This means that the measure had little or no prediction value for the majority of the trains. Due to this the regression method places almost no interest in the measure. This measure would likely prove to be more useful if we chose only peak times to predict.

4.7 Conclusion

This chapter looked at the various properties of local services in the case study of the London Fenchurch St to Shoeburyness line. It showed that there are many interesting features of train services that can be looked at and analysed. The chapter started by looking at travel times between TIPLOCs and we gave an example of how these can be used to validate timetables. It could also be used to suggest changes if necessary to allow for a more viable schedule. Extensions of this work could be looking at more schedules or devising a package to automatically check schedules and suggest change.

The next section looked at the importance of train interactions with the visualisation aid of a train graph. This method used throughout railway literature is shown as a good method to see how trains progress through the network. Concertinas were then introduced as a new way of looking at train interactions and interval graphs were used as a method of examining them. More work could be done to look at these interval graphs in more detail but the sparse nature of the graphs and the temporal nodes as well as the edges prevent more complex network structures from arising.

As a follow on to the concertina work, we looked at knock-on delay and the relationship between primary delay incidents and the propagation of this delay through concertinas to other services. Further work here would be to quantify headway as a measure of how much space ahead of a train there is to catch up to the schedule. Turn-around times were also examined and a formula for estimating turn-around times was proposed. This work could be further applied to suggest better monitoring of individual trains. We also saw how the data restricts how these can be monitored. Hopefully, a better way of tracking turn-around will be added to the data sources sometime in the future.

The last part of this chapter looked at predicting train travel times. We looked at several methods for this purpose and found that of the methods examined, the random forest model was the best. More work could be done to examine the feature selection both for the random forest and the linear regression model. Other prediction methods could be devised as well. It would be useful if such methods could be applied to the prediction model in use today on the railway. Unfortunately, that data is not published publicly so it is hard to get.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis has presented work looking at the British railway network through the lens of data-driven analysis. It has also examined how delays are propagated through interactions between trains and bottlenecks on the infrastructure. We looked at the case study of the London Fenchurch Street to Shoeburyness line to give an example of how these ideas can be applied in practice. We examined why the issues addressed in this thesis are relevant and saw how our methods could be utilised in the railway system in the future.

In chapter two we examined the berth data and constructed the national-level network of Train Describers from it and performed analysis on the network properties. The methodology was to first process and clean the data. The method to partition the data was simple but the cleaning took more work. Understanding the dataset and the railway terminology was a barrier to entry. We have used the term headcode throughout this thesis. This term has been updated to instead be train description. This is a very similar term to Train Describer, so the older term has been used instead. Many other terms are used and misused throughout the documentation available. This has made understanding the data harder than needed. The next step was to link up the signalling areas internally. This was simple because the dataset is a set of connections within areas. The sorting of trains by headcodes and then by time allowed for connections between areas to be established. This allowed the network to be ready for analysis. The code we devised and implemented was not efficient and took hours to run some of the processing steps. We could improve this coding in the future.

We then selected various network theory methods to examine the constructed

network mathematically. Centrality was the focus of our investigation of the network because this measure shows the priority of the network. The methods were implemented using the Networkx package in Python. Another metric for centrality was proposed by our industrial partners. This was train interactions divided by the number of berths in a Train Describer. This was not included in the final thesis because the results showed similar Train Describers to the other measures. We showed the various issues with the publicly available data and solutions to overcome them. Of particular note was the lack of truly unique identifiers for trains in the data set. This forced us to look at the schedule data for a solution. This solution of using trainUIDs was not perfect but allowed for a viable visualisation, as shown in figure 2.2. The subsequent centrality analysis showed that certain train describers achieved high values due to a combination of historical significance but also modern constructions like London Bridge. The main finding was that this data set confirms that the network is centred around London. But that this could be due to the lack of electrified railways elsewhere in the country. This work could be applied to other railways in the world since most use fixed-block signalling. The network in the United States would differ due to the emphasis on freight travel [96]. The work would be better compared to the European railway networks. France is considered to have a hub centred around Paris. [90]. Due to the adoption of the railways at similar times the method we developed would perform better in European countries. Japan would be a case study of interest since it is an island country of comparable GDP and population. It also has a capital with a disproportionate amount of the population.

We next looked at a finer-grain view of the network, that of the berth-level connections. The major issue seen here was the area transitions, an issue caused by historical localised thinking. This thinking was based on the idea that signalmen should only see what is in their area and no concern was placed on the overall picture. This led to poorly defined border regions that are hard to identify. We proposed a set of rules by which berths in these regions can be detected. These rules can combine multiple digital berths under one identifier thereby allowing identification of transitions. The full berth network was then analysed in a similar way to the Train Describers network. With this analysis we saw that the most central berths were concentrated in the Leamington Corridor which is a Train Describer in the Midlands. The community analysis simply grouped up the Train Describers, occasionally grouping up closely connected neighbouring TDs. The last section of the chapter looked at local solutions to berth mapping. Here we examined the London Fenchurch St to Shoeburyness line. Three issues were resolved to achieve an accu-

rate berth diagram. These were snooping berths, a relatively recent phenomenon where signallers are interested in important berths from nearby areas so monitor them. This is seen in the data to be similar to edge berths but with the four defined rules in 2.5.2, they can split correctly. The second was the edge berths which were solved at the same time as the previous problem. The last was duplicate headcodes this issue required careful handling with another rules-based solution. This finally achieved a correct answer with the real topology shown. This chapter showed that the current issues with network analysis are caused not by incorrect data entry but instead by poor long-term planning. The uptick in snooping berths shows the signallmen's interest in other areas not under their control. Whilst this can give us interesting information as to where the experts are looking, it also makes it harder to interpret the underlying topology of the real network. Network Rail need to modernise their data fields to allow for inter-area movements and abolish the snooping berths and allow signallers to access each other's data. The work done here was in consultation with our industrial partners who were interested in the underlying structure of Train Describers. The method was refined and improved until it was considered complete.

Chapter three continues the work looking at the London Fenchurch St to Shoeburyness line. We chose this line due to good data coverage and relative lack of freight as well as only one passenger operator. In hindsight, this was a good decision since even with the added simplicity there were many complications that had to be addressed. If a line without these advantages had been chosen it would have taken even more effort to allow for any analysis and it would have been harder to interpret such work. In the chapter and in appendix A we provided a comprehensive breakdown of the schedule data. This work entailed the combination of the berth data and the schedule. This did require some other data sets with one of those not available publicly. This problem could inhibit others from reproducing the work we have done here. To improve this for future researchers in the area the code used to create the appended berth data has been made available online. The link to this is at the end of appendix A.

In chapter 3.3 we looked at Sectional Running Times (SRT), these are an easy to calculate but very useful metric of use for much later work in this thesis. They are more useful if calculated with more data but the data we possessed for this work was five weeks long. This is either large or small depending on how the SRTs are split. For example, we have only five Sundays in the data set, but we do have twenty-five weekdays. More data would have been helpful for this task but would have consumed more time to process. With that balance in mind, five weeks is a

good amount. The analysis of the SRTs showed that the mean tends to be larger than the median. This was caused by the heavy tail. This tail is present because in the real-world interpretation of this SRT, a train has a speed limit so there is a lower-bound for travel time. There is no upper bound so, if a train has to wait in a berth for the track ahead to clear there will be a large travel time recorded. This affect causes the heavy tail. These SRTs were then used to form a capacity model describing how many trains can traverse a length of track in a given time. This is useful for identifying bottlenecks or, in the case shown in the chapter, identifying line features. The method behind the calculation of SRTs was chosen to include all of the available data we had. This helped allow for more robust values of travel time.

We then looked at super-threshold delay at individual TIPLOCs on a directional basis. This is plotted against the occupancy rate, which is defined as a percentage of capacity being utilised. The results seen here show how the super-threshold delay increases as trains travel down the line. Issues regarding turn-around times can be seen here also. These are covered better in chapter four. We can see the flow of traffic down the line as well with the TIPLOCs between London Fenchurch Street and Barking being the busiest. There is also lower usage between Upminster and Pitsea due to the alternative route on the branch line. A better method to detect the build-up of delay would be possible with the addition of offsets for junctions. They have entries in the schedule but without offsets they cannot be compared to the berth movements.

The last section we examined in this chapter was using a random forest regression approach to attribute delay. The implementation of this method was done using the scikit learn package in Python. Care was taken to prevent data leakage into the prediction and verification datasets. With more data this would have been easier and more robust. This was done to attempt to diagnose how significant the infrastructure is in regard to a late train arrival at a station. Permutation importance was used due to the mostly categorical data in use for the prediction. The most important feature was found to be the time of day. This is not a surprise because a significant difference in delay against time can be seen in figures 3.11-3.15. The second most important feature was the TIPLOC of arrival this and the third highest, direction of travel are both infrastructure properties. A variant of this model was also examined which constructs a random forest for each TIPLOC. The conclusion here was that the time of day was the most useful prediction metric again. Now without the TIPLOC as a feature the direction is the next most informative feature. This shows us the importance of the infrastructure when examining delays. Perfor-

mance of the model was not the focus in this chapter since the feature importance values and their real-world interpretation were the outcomes desired here.

In chapter four of this thesis, we looked at the analysis of services on a line. Again, it looks at the London Fenchurch St to Shoeburyness line, this was for the same reasons as the last chapter. Just as the last chapter examined the infrastructure this chapter looked at the train services utilising the rails. The first section was an application of the SRTs seen in chapter three. The use case here was calculating travel times between stations and using these times to validate the timetable. We also looked at how allowances work in practice to alter timetables. The method was applied to study a schedule and check each stop. This also gave us a look at stopping times at TIPLOCs. The actual dwell time is unfortunately very hard to estimate from the data that we had access to. The method we devised needs more work to examine how trains can be too close to other services and then measure the robustness of the train interactions.

Train graphs were then introduced as a way of looking at services on a distance time graph. We also explain how the branch line is visualised given that this method can only show a one-dimensional railway line. These graphs are used for the rest of this chapter to examine train interactions. These were hard to visualise and the method chosen to do so was complex. The result though was effective in showing the trains in relation to one another. One method that was attempted but was not ready for the thesis was plotting the future projections of trains for the travel times predictions in section 4.6. These would have added a useful visual aid there. The next section looked at train interactions through the examination of "Concertinas". We introduced this phenomenon and explained that it is when two trains approach close enough that the rear train must slow down to account for the train in front. The next step taken was to introduce a model that can handle the interactions between trains and record them. This would allow for accurate measurements of concertinas. The rules that we use for concertinas were defined but on reflection it would have been better if we had been able to implement a better way of trains entering concertinas. The four-berth entry rule allows concertina tracking to be smooth, but this comes at the cost of some improperly recorded concertinas. Therefore, there are slightly more concertinas observed than would be ideal in our model when compared with a three-berth entry rule. A short section then follows which discusses concertina chains and models them as interval graphs. These chains are plotted and then we examined how they grow and shrink. We then show an altered train graph that has been modified to display when trains are in concertina.

Knock-on delay is examined in the next section, and we again used the train

graphs to show how delay can propagate between services. These delays are caused by the concertinas that we looked at in the last section. We took an in-depth look at a particularly disruptive day of data. This examination then led to our discovery of turn-around delays. The next step we looked at was to examine these turnarounds in more detail. The turn-around estimation formula was shown as a way of calculating the delay a service would depart with given how late it arrived and how much leeway there was to turn around with. We did not however look at testing this proposed formula to see how it performs against unseen data. A few more train graphs are shown, these are less disruptive examples of schedules being carried out with less incidents. The graphs are compared with the metric of delay per train. This allows for a quantifiable comparison between different days of data that the graphs are unable to convey.

The last section we looked at in this chapter was the prediction of arrivals at TIPLOCs. We observed how the current state of the network as well as historical data could be used to predict when a train would arrive at a TIPLOC in the near future. The chosen baseline was the timetable and this was beaten by the methods chosen. The methods we looked at were the two SRT methods that used current location and movements from the previous week to add Sectional Running Times to the current time of day and then this answer would be compared to the actual time of arrival. This was the obvious method because it utilised previously calculated values and allowed for a simple method that would outperform the schedule. The difference between the two SRT methods was that the second method uses SRTs split by type of train, e.g., freight, parcel, passenger. A linear regression method was also proposed and was given the baseline of the timetable and both the SRT methods. It also received several other input values such as time of day and the day of the week that the prediction occurred on. These are the same features used at the end of chapter three. Feature selection was then used to improve the prediction. The method used here was to add or remove features and evaluate the performance of the method. Even with this the linear regression method was not much better than the SRT one. The last method used was a random forest regression model which was given all eight of the features that were given to the linear regression. Feature selection was also used here to improve how the method performed. Fortunately, the random forest regression method performs better than the rest in all metrics examined. The permutation importance shows that the schedule data and the SRT by class inputs are the most useful features. This could show that some of the other features are not useful for general predictions and could be more helpful for specific cases. Our random forest had an average absolute median error of 10.67 seconds

and an average absolute mean error of 21.41 seconds. This is based on predicting randomly at a distance of between five and ten berths away. Therefore, these are short-term predictions so would perform worse over longer distances.

5.2 Future Work

There is much more work that could be done on the study areas examined in this thesis. In chapter two more work could be done to investigate and validate each connection between Train Describers to look for suspicious movements. If these could be found and filtered out, we could have more confidence that the connections that are seen are legitimate. We could also look at improved accept/reject criteria for rural lines to allow for North Wales to be attached to the rest of the network. Other analysis could also be performed to add weight to the edges to match corresponding traffic quantities. For the visualisation of the British berth network, we could use a different display algorithm that would take longer to calibrate for a clearer picture. Some of the later methods discussed in this thesis could also improve the edges seen here. We are satisfied that the berth network for the local line is almost perfect. However, a better job could be done to display stations and their respective platforms.

More work could also be done in chapter three to investigate the various natures of the SRTs. They could be classified for their various roles e.g., platform berth, siding berth, edge berth, snooping berth. Each of these types could be examined in isolation from one another and would likely show different distributions as well. In regard to bottleneck analysis we could look at each platform in turn to see if each of them is being used to the maximum available. Also, if data could be found for dwell times that would be of real value to see how they change at peak times. The random forest method seen at the end of the chapter could be improved by adding more features and having more data to input into the model. Both actions could be done with more time.

Chapter four could be furthered in scope as many of the sections had more work that could have been done. The timetable validation in section 4.2 could be automated and used to check all of the timetables for plausibility and possibly use SRTs tailored to the time of day being looked at. The train graphs could be improved if we could devise a way to handle multiple berth paths for larger stations because at the moment these platforms are not handled correctly. Further work could be done to look at the interval graphs generated by the concertinas. Unfortunately, these graphs are exceedingly sparse and have both temporal nodes and temporal edges

causing harder than normal analysis. The study of concertinas has applications to all berth-based rail systems including light-rail (trams). Future studies could be done to examine concertinas application to train braking and look at the loss of fuel efficiency caused by repeated acceleration. We would be interested in studying large delays further as well. This could be extended into primary delay analysis if we had access to the data from the Delay Attribution Board. Further testing could be done on the turn-around formula to check it on unseen data. The prediction models at the end of the chapter could look at different TIPLOCs and maybe extended to different Train Describers of the Network Rail system. We could also examine how services managed by different operators interact with one another.

Earlier in our research we looked at a simulated network model that could examine various situations that could be created and examined. Part of the method behind this model went into the model used to locate concertinas in chapter four. It was not intended for this purpose however and we were unable to continue this work as we ran out of time. This would be of interest to explore at a later time.

The model worked by examining basic network topologies to test delays when given different schedules. The below figure shows a topology that causes conflicts.

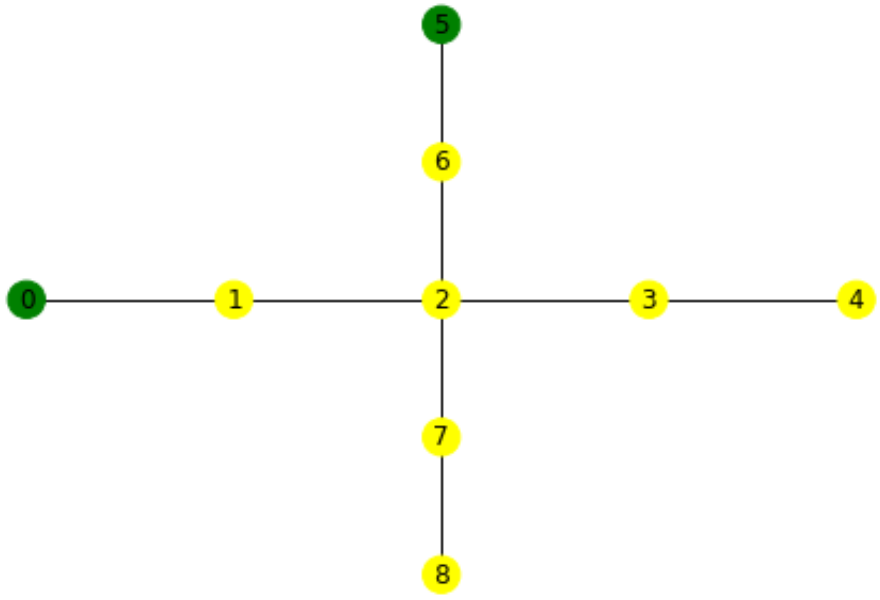

Figure 5.1: Simple topology showing a conflict point.

The green nodes in figure 5.1 show origin berths or stations. Node two is the

conflict node since trains from both origin points travel through the shared node. This is a simple model that can examine train-on-train delays. The advantages of this model are that the length of track segments is known and the schedule can be changed without causing disruption. The capacity of the nodes can be changed to model an increased number of platforms. Delay can be measured and schematics can be changed to affect it.

The model works through a series of look-up tables. These store various attributes of the network such as the current state of the trains, the berths and the schedule. The last dataframe stores the path which the trains take through the network. It is this concept that was adapted for use in the concertina model in chapter four.

Different topologies were looked at such as, a straight line of five berths, the graph shown in figure 5.1 and a more complex cross pattern that was designed to examine indirect delay. Ultimately, there was no significant output from this work. It was planned to use real network topologies in the future to act as a digital twin. This could be worked on in the future. The original research plan for an additional chapter to this thesis was to adapt this toy model for use on the berth network for the London Fenchurch St to Shoeburyness line. A rules-based system was designed to prevent conflicts and enable different schedules. The reason behind not doing this was that no method could be devised and implemented to document or replicate primary delays. But with data from the Delay Attribution Board (DAB) this issue could be solved in the future. We are confident that secondary delays could be accurately modelled within the existing toy model. Turn-around times were also a concern but if solved they would allow for greater time horizons for predictions in chapter five of this thesis also.

A queuing theory model was also looked at for simple topologies on the toy models. The general theory would be that trains/services would be the customers and berths/stations would be the servers. Queuing theory could be used to examine issues at an individual station level[10]. Applying the theory to a whole line was examined but quickly it was found that too many simplifications would be required to allow for the model to operate. In Bychkov's paper [10] a queuing network was used for a single station. Therefore, a network of queuing networks would be required to accurately model a whole line.

The lack of a model for the line in this thesis means that alternate scenarios cannot be looked at. This prevents us from adding platforms to stations in the model and examining the potential impact of such a move. This leaves us with data analysis as the only option to look at the line.

Appendix A

Processing of the berth and schedule data

A.1 Breakdown of timetable data

A.1.1 Schedule data description

This section will address the schedule data that was mentioned in chapter 2.3. This schedule data set is also available from Network Rail's data feeds. The start of the dataset looks like this:

```
AANC03858Y0303919051919090800000001NPSNWCSTLE  TO                                P
AANC09086C0908319051919120800000001NPSPARKGAT  TO                                P
AANC09092C0908919051919120800000001NPSPARKGAT  TO                                P
AANC09098C0909519051919120800000001NPSPARKGAT  TO                                P
AANC09104C0910119051919120800000001NPSPARKGAT  TO                                P
```

These are all "AA" type messages that describe train associations and we will not be using this data. The more useful information is located further down the dataset. Below is shown the schedule for a single service:

```
BSNC001481905191912080000001 PEE5P01    124650005 EMU    090D                 P
BX         SEY
LOGLNGDEP 0142 0000          TB
LIGLNGREC 0152 0153      00000000          OPC
LTGLNGHMK 0155 00001     TF
```

This schedule shows five different message types that contain different information, and each type will be addressed in turn. The five types are BS, BX, LO, LI and LT. They respectively stand for:

1. BS - Basic Schedule (one per service) 2. BX - Basic schedule eXtra details (one per service) 3. LI - Location Origin (one per service) 4. LO - Location Intermediate (one for each intermediate stop for the service) 5. LT - Location Terminus (one per service)

The first line is a "BS" message which stands for basic schedule. It contains all the relevant information required about the service other than the locations it will stop at which are described in the three "L" type messages.

Message Type	Transaction Type	TrainUID	Date From	Date To	Days Run	Bank Holiday	Train Status	Train Category	Train identity	Headcode	Power Type	Service code	Speed	Operate Character
BS	N	C00148	190519	191208	0000001		P	EE	5P01		EMU	2465005	090	D

Figure A.1: Breakdown of BS data from the Network Rail datafeeds.

The first column specifies that this is indeed a basic schedule message. The second column highlights the type of transaction, in this case it is N which stands for new. The next column has the train UID which was mentioned in chapter 2.3. The date from and date to column show when the train runs. This example has the service operating for three months of the timetable, starting on the 19th of May 2019 and finishing on the 6th of August 2019.

The days run column shows on which days of the week this service runs with a 1 if the service operates or 0 if not. This service runs only on Sundays. The next column is used to discuss if the service runs on bank holidays, if left blank it does not, which occurs in this case. The train status column contains the service type which can be passenger train, freight train, bus or even ship. There are also temporary changes for replacements where necessary.

The train category better describes the type of train service. This service is of type "EE", meaning that it is empty coaching stock. The train identity column contains another name for the train outside of the train UID. Train identity is another name for the headcode of the train. In earlier data sets the headcode is not blank and instead contains the same value as the train identity column. The power type specifies how the train is fuelled, in this case EMU is Electrical Multiple Unit. The service code is used to divide trains into different service groups and is typically used for attribution of revenue. The speed column shows the planned maximum speed of the service in miles per hour. The operating character column contains up to six different characters describing other aspects of the train. The example given

here contains only “D” informing the receiver that this train contains just the driver and no passengers. This is another indicator that this train is empty coaching stock. There are other categories in the data which have been omitted here, since they are mostly left blank or always contain the same entry, for the sake of brevity.

The “BX” stands for basic schedule extra details. A description of what it contains can be found in figure A.2.

Message Type	Traction Class	UIC code	ATOC code	Applicable timetable code
BX			SE	Y

Figure A.2: Breakdown of BX entry from the Network Rail datafeeds.

Most of this message is left blank on a regular basis. Of the fields with values the ATOC code stands for the Association of Train Operating Companies. It states the operator of this particular service is SE or rather Southeastern. The applicable timetable code shows whether the train will be monitored for performance purposes or not. Y is for yes and N is for no.

The next type of message is an “LO” message, LO standing for origin location. It is the first entry in the timetable for a service. More information can be found in figure A.3.

Message Type	Location	Scheduled Departure	Public Departure	Platform	Line	Path	Engineering allowance	Pathing allowance	Activity	Performance allowance
LO	GLNGDEP	0142	0000						TB	

Figure A.3: Breakdown of LO entry from the Network Rail datafeeds.

The location column shows the TIPLOC (Timing point location) that the train starts at. GLNGDEP is the Gillingham (Kent) Electric Multiple Unit Depot. It can also have a number after the TIPLOC if that location has been visited multiple times in the schedule. The scheduled departure shows the time at which the train will depart. It can also have a "H" afterwards showing half minutes so 0142H would equate to 01:42:30 in digital format. There is no public departure time for this train because it is empty coaching stock, consequently the value returned is a dummy value of "0000". The public departure time also cannot report in half minutes. The platform column describes which platform the service set off from. Since this train originated from a depot this column is left blank. Line shows which line is planned to be used by the train at this TIPLOC and uses a three-character abbreviation of the line. Allowances are additional minutes added to scheduled travel times due to anticipated disturbances. They vary by type and allow for problems that may delay a service.

The last column is for activity which describes what happens at this TIPLOC. "TB" stands for train begins.

LI messages are Intermediate Location records. They are used for TIPLOCs between the origin and destination.

Message Type	Location	Scheduled Arrival	Scheduled Departure	Scheduled Pass	Public Arrival	Public Departure	Platform	Line	Path	Activity	Engineering allowance	Pathing allowance	Performance allowance
LI	GLNGREC	0152	0153		0000	0000				OPC			

Figure A.4: Breakdown of LI data from the Network Rail datafeeds.

Many of the fields in LI type messages are the same as in LO. There are additional columns for scheduled arrival and public arrival these follow the same rules as the scheduled departure and public departure from the LO data. There is also a scheduled pass column that has the same format as the scheduled departure but only has a value when the train does not stop. There is no public pass since the public have no need of such information. Incidentally the activity described here "OPC" refers to stopping for operating reasons "OP" and stops to change trainmen "C". So, the train has stopped to change the driver.

The last information from this data set is the LT type messages that are terminating location records. An LT record is issued for the end of the service. All services have one LO and one LT message.

Message Type	Location	Schedule d arrival	Public arrival	Platform	Path	Activity
LT	GLNGHMK	0155	0000	1		TF

Figure A.5: Breakdown of LT data from the Network Rail datafeeds.

Similarly to the LI type entry the LT message has a lot of the same columns as the LO message. The "TF" activity here stands for train finishes.

A.1.2 Compiling the timetable data

Unfortunately, the timetable needs further work before being usable. The timetable is released from Network Rail on the data feeds at the start of each month in an approximately 250MB file. This file contains 2.7 million lines of string data. Then the data must be updated each day with issued changes with data sizes of between 0 and 25MB to remain correct for further use. Ideally, this data would include the changes already applied and this could be a future improvement to the dataset. These changes can be temporary or permanent and there is a different amount each day to address the changes being made. We use data for five weeks in this thesis, from September 1st, 2019, to October 6th, 2019. Fortunately, the October dates were already processed for us. This does leave the month of September to be worked

on.

This combining is done mostly with the use of STP indicators. They are another field contained within BS messages at the start of a train record. There are four types of these messages each having a different purpose. The first is P type messages these represent permanent additions to the timetable for all future days. The next is N type messages which are temporary additions to the timetable these are typically added for only a few days. O type messages are overlays of the existing timetable which can change trains already present to having different stopping patterns or arrival times or even possibly a new train type. The last type of STP message are C-type messages. They differ from the rest of the indicators since they have no LO, LI or LT entries associated with them. These are cancellation messages which remove services from the schedule for the specified period of time.

To create a month of schedules from this data we first initialised thirty separate files that could each be altered for the data relevant for that day. They are all initially composed of the P-type messages from the 1st of September data set containing all trains for that month. These different data sets can then be read into the code, altered and then written out again. Each message type is applied in turn matching on the basis of the trainUID which allows the correct services to be removed or altered.

For each day in the data set all of the types of messages are processed in turn. The P type records are the easiest since they apply to all days in the data set after and including the date that the P record was issued. So, all the relevant days then have this record appended to them. The C-type messages cover certain days, so those days are then read into the code and the relevant trainUIDs are matched allowing for the removal of these records.

The N type records are similar to the P types, but they are only applied to the days specified in the period covered by the Date From - Date To entries. O type records are the most complex because the previous record to be altered must be removed using the same method as is used in the C records and then the new version is added in a similar way to the N record type.

This method is applied sequentially to all the days in the month, which in this case is thirty. The finishing days of the data are now the compiled timetable, ready for further work.

A.1.3 TIPLOC to STANOX data

Another data set that will be used int this chapter is the TIPLOC to STANOX (station number) reference list. Historically TIPLOCs were used for the passenger

systems and STANOXs were for freight services. The systems now exist alongside one another, and they can be matched to some extent. The mapping between the two is a many-to-many relationship with many of the differences being in freight depots.

This body of work is focusing on a commuter passenger line so this difference is not significant. The data set originates from a publicly available website [18] and a sample is shown in figure A.6.

Location	CRS	NLC	TIPLOC	STANME	STANOX
Aachen	NaN	81601	AACHEN	AACHEN	00005

Figure A.6: Breakdown of the TIPLOC-STANOX dataset [18].

There are almost fifteen thousand such entries in this data set. Some of which have multiple entries in the STANOX column that cause parsing errors in the data frame, these were removed by hand due to the infrequent existence and the format we received the data in. The location column simply gives a real-world place that is located near this section of the network. The CRS code (Computer reservation system) is a way of identifying locations and was created for the seat reservation system [18]. NLC codes (National location codes) are used for asset identification and are used for financial purposes [18]. We have already discussed the inclusion of the TIPLOC column. The STANME code (Station number names) describes a STANOX in nine characters, but it can be longer in some cases [18]. Lastly STANOXs are five-digit codes that are mostly unique that describe locations on the network [18]. This data set has large amounts of missing values but for the purposes of our work is sufficient.

A.1.4 STANOX to berth data set

The last data set used in this chapter is the STANOX to berth data. We obtained this data set from our industrial partners which they, in turn, received from one of their customers. A sample of this data is shown below:

Train Describer	Step Type	From Berth	To Berth	STANOX	Platform	Passing Offset	Stopping Offset
IH	I	NaN	7904	01125	1	0	0
IH	I	NaN	7906	01125	2	0	0
IH	I	7903	7909	01125	1	0	0
IH	B	7904	7909	01125	1	25	25

Figure A.7: Breakdown of the STANOX-berth dataset.

This relates STANOXs to "trigger berth pairs", which are berths that are utilised for the purpose of recording trains arriving at locations. This data set focuses on stations mainly and lacks trigger berths for junctions, this affects our later analysis.

The first column says which train describer the berths are in, "IH" is Inverness Highlands west CAD. Column two is the step type, "I" is an interpose meaning the train is entering the network. "B" is for a standard movement on the line from one berth to another. From berth and to berth both have to be matched for the berth movement to be linked to the location. The next column shows the STANOX name for the location. The platform column shows which platform at the station the berths correspond to. If this is a junction a NaN value will be in the column instead.

The two offset columns show the value in seconds that is added to the berth movement time to approximate arrival at the location. Passing offset if the train does not stop and the stopping offset otherwise. In practice however the data set has the same values in both of these columns 99.84 percent of the time. This shows that the data set is far from complete since in the cases where these values differ, they do so significantly. Differences can be as great as six minutes in some cases.

These datasets are combined in section 3.2 of the main text of the thesis. The code for this data processing can be found at https://github.com/AndrewHilditch/Data-Cleaning/tree/develop.

Acronyms

BS Basic schedule. 90

BX Basic schedule extra details. 90

C2C Coast to coast. 24

CA Berth step. 11

CB Berth cancel. 11

CC Berth interpose. 11

CRS Computer reservation system. 95

DAB Delay attribution board. 41

ES East Sussex. 15

GB Great Britain. 10

LI Intermediate location. 90

LO Origin location. 90

LT Terminus location. 90

NLC National location codes. 95

SF Signalling update. 11

SRT Sectional running time. 38

STANME Station number names. 95

STANOX Station number. 94

STP Short term planning. 94

TIPLOC Timing point location. 92

TOC Train operating company. 9

trainUID train unique identifier. 12

ZE Eastbourne. 15

Bibliography

[1] Mona Al Hamad and Ahmed M Zeki. Accuracy vs. cost in decision trees: A survey. In *2018 international conference on innovation and intelligence for informatics, computing, and technologies (3ICT)*, pages 1–4. IEEE, 2018.

[2] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[3] John Armstrong and John Preston. Capacity utilisation and performance at railway stations. *Journal of rail transport planning & management*, 7(3):187–205, 2017.

[4] Mehmet Şirin Artan and İsmail Şahin. Exploring patterns of train delay evolution and timetable robustness. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[5] Ahmad Taher Azar and Shereen M El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7):2387–2403, 2013.

[6] Aveek Bhattacharya. Out of business? *IPPR Progressive Review*, 27(1):58–68, 2020.

[7] Phillip Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007.

[8] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[9] Tom Burridge. Hs2: Give me the facts, says shapps, Jan 2020.

[10] Igor Bychkov, Alexander Kazakov, Anna Lempert, and Maxim Zharkov. Modeling of railway stations based on queuing networks. *Applied Sciences*, 11(5):2425, 2021.

[11] Mark Casson. The future of the uk railway system: Michael brookes vision. *International Business Review*, 13(2):181–214, 2004.

[12] Lei Chen, Clive Roberts, Felix Schmid, and Edward Stewart. Modeling and solving real-time train rescheduling problems in railway bottleneck sections. *IEEE Transactions on intelligent transportation systems*, 16(4):1896–1904, 2015.

[13] Coast 2 Coast, Sep 2016.

[14] Piers Connor. British signalling what the driver sees v2 - railway technical, 2017.

[15] Pierluigi Coppola and Francesco De Fabiis. Impacts of interpersonal distancing on-board trains during the covid-19 emergency. *European Transport Research Review*, 13(1):1–12, 2021.

[16] Inês Costa-Carrapiço, Rokia Raslan, and Javier Neila González. A systematic review of genetic algorithm-based multi-objective optimisation for building retrofitting strategies towards energy efficiency. *Energy and Buildings*, 210:109690, 2020.

[17] Adele Cutler, D Richard Cutler, and John R Stevens. Random forests. In *Ensemble machine learning*, pages 157–175. Springer, 2012.

[18] Phil Deaves. Crs, nlc, tiploc and stanox codes, Aug 2021.

[19] Marko Debeljak and Sašo Džeroski. Decision trees in ecological modelling. In *Modelling complex ecological dynamics*, pages 197–209. Springer, 2011.

[20] Ross Dentten. Rail industry focus, Jan 2015.

[21] Rui Ding, Norsidah Ujang, Hussain bin Hamid, and Jianjun Wu. Complex network theory applied to the growth of kuala lumpurs public urban rail transit network. *PloS one*, 10(10):e0139961, 2015.

[22] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

[23] Andrew Disney. Pagerank centrality and eigencentrality, Jun 2021.

[24] Alexander Doll, Dirk Friebel, Matthias Rückriegel, and Christian Schwarzmüller. Global logistics markets. *Munich: Roland Berger Strategy Consultants*, 2014.

[25] Bryan Donnelly. pdf rdg/ec/gn/005 on-train satellite navigation equipment/darwin interface (issue 2), Apr 2018.

[26] César Ducruet and Theo Notteboom. The worldwide maritime network of container shipping: spatial structure and regional dynamics. *Global networks*, 12(3):395–423, 2012.

[27] Helene Dyrhauge. Diverging national railway policies. In *EU Railway Policy-Making*, pages 15–30. Springer, 2013.

[28] Andrea DAriano, Dario Pacciarelli, and Marco Pranzo. Assessment of flexible timetables in real-time traffic management of a railway bottleneck. *Transportation Research Part C: Emerging Technologies*, 16(2):232–245, 2008.

[29] Xiao Feng, Shi-Wei He, and Yu-Bin Li. Temporal characteristics and reliability analysis of railway transportation networks. *Transportmetrica A: Transport Science*, 15(2):1825–1847, 2019.

[30] Garrett Fullerton, Giovanni C DiDomenico, Mei-Cheng Shih, and C Tyler Dick. Congestion as a source of variation in passenger and freight railway fuel efficiency. In *ASME/IEEE Joint Rail Conference*, volume 45356, page V001T07A002. American Society of Mechanical Engineers, 2014.

[31] Eliane A Gattass and George L Nemhauser. An application of vertex packing to data analysis in the evaluation of pavement deterioration. *Operations Research Letters*, 1(1):13–17, 1981.

[32] Robert Geisberger, Peter Sanders, and Dominik Schultes. Better approximation of betweenness centrality. In *2008 Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 90–100. SIAM, 2008.

[33] Stephen Gibson, Grahame Cooper, and Brian Ball. Developments in transport policy: The evolution of capacity charges on the uk rail network. *Journal of Transport Economics and Policy (JTEP)*, 36(2):341–354, 2002.

[34] Martin Charles Golumbic. Interval graphs and related topics. *Discrete Mathematics*, 55(2):113–121, 1985.

[35] Michael F Gorman. Statistical estimation of railroad congestion delay. *Transportation Research Part E: Logistics and Transportation Review*, 45(3):446–456, 2009.

[36] Rob MP Goverde. Punctuality of railway operations and timetable stability analysis. 2005.

[37] Rob MP Goverde. Railway timetable stability analysis using max-plus system theory. *Transportation Research Part B: Methodological*, 41(2):179–201, 2007.

[38] Ulrike Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

[39] Michel Habib, Christophe Paul, and Mathieu Raffinot. *Common connected components of interval graphs*. PhD thesis, LIRMM (UM, CNRS), 2003.

[40] Naohiko Hibino, Hisao Uchiyama, and Yoshihisa Yamashita. A study on evaluation of level of railway services in tokyo metropolitan area based on railway network assignment analysis. *Journal of the Eastern Asia Society for Transportation Studies*, 6:342–355, 2005.

[41] Peter Hicks. Live map of london fenchurch street to shoeburyness, Mar 2015.

[42] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[43] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[44] Nantian Huang, Guobo Lu, and Dianguo Xu. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies*, 9(10):767, 2016.

[45] Ping Huang, Chao Wen, Liping Fu, Qiyuan Peng, and Zhongcan Li. A hybrid model to improve the train running time prediction ability during high-speed railway disruptions. *Safety Science*, 122:104510, 2020.

[46] Alexandre Jacquillat and Amedeo R Odoni. An integrated scheduling and operations approach to airport congestion mitigation. *Operations Research*, 63(6):1390–1410, 2015.

[47] Chaozhe Jiang, Ping Huang, Javad Lessan, Liping Fu, and Chao Wen. Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression. *Canadian Journal of Civil Engineering*, 46(5):353–363, 2019.

[48] Clive Kessell. Single line internet control, Jun 2019.

[49] Clive Kessell. Delivering better timetables, Apr 2020.

[50] Fahimeh Khoshniyat and Anders Peterson. Improving train service reliability by applying an effective timetable robustness strategy. *Journal of Intelligent Transportation Systems*, 21(6):525–543, 2017.

[51] Kap Hwan Kim and Hoon Lee. Container terminal operation: current trends and future challenges. *Handbook of ocean container transport logistics*, pages 43–73, 2015.

[52] Florian Knorr, Daniel Baselt, Michael Schreckenberg, and Martin Mauve. Reducing traffic jams via vanets. *IEEE Transactions on Vehicular Technology*, 61(8):3490–3498, 2012.

[53] Dmytro Kozachenko, Vladyslav Skalozub, Bogdan Gera, Yuliia Hermaniuk, Ruslana Korobiova, and Aleksandra Gorbova. A model of transit freight distribution on a railway network. *Transport problems*, 14, 2019.

[54] Shalaka Kurup, David Golightly, David Clarke, and Sarah Sharples. Passenger information provision: Perspectives from rail industry stakeholders in great britain. *Journal of Rail Transport Planning & Management*, 19:100264, 2021.

[55] Chi Yung Lam and Kang Tai. Network topological approach to modeling accident causations and characteristics: Analysis of railway incidents in japan. *Reliability Engineering & System Safety*, 193:106626, 2020.

[56] Alex Landex. Capacity statement for railways. In *Selected Proceedings from the Annual Transport Conference at Aalborg University*, volume 2, 2007.

[57] Alex Landex and Anders H Kaas. Planning the most suitable travel speed for high frequency railway lines. In *Proceedings of the 1st international seminar on railway operations modelling and analysis*, pages 1–16, 2005.

[58] Florent LAROCHE. An exploration of railway congestion: A proposed method for estimating the infrastructure utilization rate.

[59] Tae-Hwy Lee, Aman Ullah, and Ran Wang. Bootstrap aggregating and random forest. In *Macroeconomic forecasting in the era of big data*, pages 389–429. Springer, 2020.

[60] Wei-Hsun Lee, Li-Hsien Yen, and Chien-Ming Chou. A delay root cause discovery and timetable adjustment model for enhancing the punctuality of railway services. *Transportation Research Part C: Emerging Technologies*, 73:49–64, 2016.

[61] Ross Liddell. The response of south western railway to the covid-19 pandemic. *Journal of Risk Analysis and Crisis Response*, 11(2):57–66, 2021.

[62] Ryan Lilley. Rdg delay attribution review report 2020-09-28, May 2020.

[63] Anders Lindfeldt. *Railway capacity analysis: Methods for simulation and evaluation of timetables, delays and infrastructure*. PhD thesis, KTH Royal Institute of Technology, 2015.

[64] Jintao Liu, Felix Schmid, Wei Zheng, and Jiebei Zhu. Understanding railway operational accidents using network theory. *Reliability Engineering & System Safety*, 189:218–231, 2019.

[65] Alistair I Mees. Railway scheduling by network optimization. *Mathematical and Computer Modelling*, 15(1):33–42, 1991.

[66] Juan Montero and Matthias Finger. Railway regulation: a comparative analysis of a diverging reality. In *Handbook on Railway Regulation*. Edward Elgar Publishing, 2020.

[67] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[68] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

[69] Yasufumi Ochiai, Yoshiki Masuma, and Norio Tomii. Improvement of timetable robustness by analysis of drivers' operation based on decision trees. *Journal of Rail Transport Planning & Management*, 9:57–65, 2019.

[70] Koji ONO, Naotugu NOZUE, Masami KOTANI, and Yuuko ABE. A system for simulating train transportation and an analysis of congestion. *Quarterly Report of RTRI*, 40(4):219–222, 1999.

[71] Alessio Pagani, Guillem Mosquera, Aseel Alturki, Samuel Johnson, Stephen Jarvis, Alan Wilson, Weisi Guo, and Liz Varga. Resilience or robustness: identifying topological vulnerabilities in rail networks. *Royal Society open science*, 6(2):181301, 2019.

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[73] Suporn Pongnumkul, Thanakij Pechprasarn, Narin Kunaseth, and Kornchawal Chaipah. Improving arrival time prediction of thailand's passenger trains using historical travel times. In *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 307–312. IEEE, 2014.

[74] Peter Pudney and Alex Wardrop. Generating train plans with problem space search. In *Computer-aided systems in public transport*, pages 195–207. Springer, 2008.

[75] Network Rail. Data feeds, Oct 2012.

[76] Network Rail. About us, Feb 2021.

[77] Saul G Ramirez, Riley Chad Hales, Gustavious P Williams, and Norman L Jones. Extending sc-pdsi-pm with neural network regression using gldas data and permutation feature importance. *Environmental Modelling & Software*, 157:105475, 2022.

[78] A Robinson and A Van Niekerk. Uses of anpr data in traffic management and transport modelling. 2014.

[79] Andrew J Sage, Yang Liu, and Joe Sato. From black box to shining spotlight: Using random forest prediction intervals to illuminate the impact of assumptions in linear regression. *The American Statistician*, (just-accepted):1–26, 2022.

[80] Mark R Segal. Machine learning benchmarks and random forest regression. 2004.

[81] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

[82] Peter Sels, Thijs Dewilde, Dirk Cattrysse, and Pieter Vansteenwegen. Reducing the passenger travel time in practice by the automated construction of a robust railway timetable. *Transportation Research Part B: Methodological*, 84:124–156, 2016.

[83] Alexander N Shabelnikov and Ivan A Olgeyzer. Technology and mathematical basis of digital twin creation in railway infrastructure. In *International Conference on Intelligent Information Technologies for Industry*, pages 688–695. Springer, 2019.

[84] Masoud Shakibayifar, Erfan Hassannayebi, Hamid Mirzahossein, Shaghayegh Zohrabnia, and Ali Shahabi. An integrated train scheduling and infrastructure development model in railway networks. *Scientia Iranica*, 24(6):3409–3422, 2017.

[85] Jay Simmonds. Estimates of station usage 2019-20, Dec 2020.

[86] Marina Skurichina and Robert PW Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.

[87] Leo Spizzirri. Justification and application of eigenvector centrality. *Algebra in Geography: Eigenvectors of Network*, 2011.

[88] Robert Stahlbock and Stefan Voß. Operations research at container terminals: a literature update. *OR spectrum*, 30(1):1–52, 2008.

[89] H Teshima, S Hori, A Shimura, and N Sato. Railway track layout modelling and its application to an automatic route setting system. *Computers in Railways XIV. Railway Engineering Design and Optimization*, 2014.

[90] Ian B Thompson. Anew kind of location decisionwhere to build high-speed railway stations: The french case. *Scottish Geographical Magazine*, 109(2):106–110, 1993.

[91] Zhao Tian, Limin Jia, Honghui Dong, Fei Su, and Zundong Zhang. Analysis of urban road traffic network based on complex network. *Procedia engineering*, 137:537–546, 2016.

[92] A Tikhomirov, A Rossodivita, N Kinash, A Trufanov, and O Berestneva. General topologic environment of the russian railway network. In *Journal of Physics: Conference Series*, volume 803, page 012165. IOP Publishing, 2017.

[93] Gwyn Topham. Hs2 rail leg to leeds scrapped, grant shapps confirms, Nov 2021.

[94] High Speed Two. About us, Jul 2022.

[95] David Tyrrall. The uk railway privatisation: failing to succeed? *Economic Affairs*, 24(3):32–38, 2004.

[96] Jose Manuel Vassallo and Mark Fagan. Nature or nurture: why do railroads carry greater freight share in the united states than in europe? *Transportation*, 34(2):177–193, 2007.

[97] Roger Vickerman. Will covid-19 put the public back in public transport? a uk perspective. *Transport Policy*, 103:95–102, 2021.

[98] Chunan Wang and Xiaoyu Wang. Airport congestion delays and airline networks. *Transportation Research Part E: Logistics and Transportation Review*, 122:328–349, 2019.

[99] Phil Wieland. List of train describers, Sep 2015.

[100] Paul Wright. *The cognitive responses to UK railway signals during train driving.* PhD thesis, University of Reading, 2017.

[101] Masoud Yaghini, Mohammad M Khoshraftar, and Masoud Seyedabadi. Railway passenger train delay prediction via neural network model. *Journal of advanced transportation*, 47(3):355–368, 2013.

[102] Tatsuki Yamauchi, Mizuyo Takamatsu, and Shinji Imahori. Optimizing train stopping patterns for congestion management. In *17th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[103] Jianxin Yuan and Ingo A Hansen. Closed form expressions of optimal buffer times between scheduled trains at railway bottlenecks. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 675–680. IEEE, 2008.

[104] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press, 2014.

[105] Yunshi Zhao, Julian Stow, and Chris Harrison. Improving the understanding of spad risks using red aspect approach data. In *Safety and Reliability*, volume 36, pages 199–212. Taylor & Francis, 2016.

[106] Yunshi Zhao, Julian Stow, and Chris Harrison. A method for classifying red signal approaches using train operational data. *Safety science*, 110:67–74, 2018.